

Robust Low-Complexity Randomized Methods for Locating Outliers in Large Matrices

Xingguo Li, *Student Member, IEEE*, and Jarvis Haupt, *Member, IEEE*

Abstract—This paper examines the problem of locating outlier columns in a large, otherwise low-rank matrix, in settings where the data are noisy, or where the overall matrix has missing elements. We propose a randomized two-step inference framework, and establish sufficient conditions on the required sample complexities under which these methods succeed (with high probability) in accurately locating the outliers for each task. Comprehensive numerical experimental results are provided to verify the theoretical bounds and demonstrate the computational efficiency of the proposed algorithm.

Index Terms—Adaptive sensing, collaborative filtering, compressed sensing, robust PCA, sparse inference

I. INTRODUCTION

In this paper we examine a robust outlier identification problem. Given a data matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, we assume that \mathbf{M} is approximately low-rank, corrupted by (nominally few) outlier columns. More formally, we suppose that

$$\mathbf{M} \approx \mathbf{L} + \mathbf{C}, \quad (1)$$

where \mathbf{L} is a rank- r matrix and \mathbf{C} is a column-sparse matrix with k nonzero columns that are interpreted as “outliers” of the subspace spanned by columns of \mathbf{L} . Our specific goal is to identify the locations of the nonzero columns of \mathbf{C} , *without* necessarily identifying the inliers (or the subspace they span), and our particular interest in this work is in doing so when our observations of \mathbf{M} may be contaminated by additive noise, or when only a subset of elements of \mathbf{M} are available, and n_1, n_2 are possibly very large relative to the rank r and the number of outliers k .

Our investigation is motivated by a wide class of “big data” applications where the outliers themselves are of interest, such as when identifying malicious responses in collaborative filtering applications [1] or finding anomalous patterns in network traffic [2]. Another example arises in computer vision tasks where the aim is to estimation visually salient regions of images [3]–[5]; recent efforts have shown that column outlier models can be viable for describing salient image regions at the “patch” level [6], making the outlier identification approach germane to saliency map estimation tasks.

Within the context of these so-called robust principal component analysis (PCA) tasks, a number of contemporary methods have been developed, which exploit low-dimensional models within the context of convex inference methods. For example, [7], [8] examine robust PCA problems based on

entry-wise sparse corruptions, while [9]–[14] propose methods applicable when outliers are present as entire columns. Despite their provable analytical successes, these methods can be computationally demanding when applied to very large data matrices, and more notably for our purposes here, these existing techniques seek to identify or approximate the low-rank matrix \mathbf{L} or the subspace spanned by its columns. Here, our interest is only in locating the outlier columns, and we seek inference procedures having both low sample and implementation complexities (e.g., to obviate the need to store and process the full data matrix).

A. Overview of Our Contribution

Our initial investigation along these lines [15], [16] proposed a randomized two-step procedure, called adaptive compressive outlier sensing (ACOS), for locating column outliers of a matrix $\mathbf{M} = \mathbf{L} + \mathbf{C}$ (i.e., in a noise-free setting, where all matrix elements are available). The key innovations associated with this approach were the utilization of dimensionality reduction methods, along the lines of those employed in compressed sensing and related areas [17], [18], and sequential adaptive sensing methods, motivated by [19]–[25], where sampling actions are allowed to depend on previous measurements. In our prior work, we showed that when $k = \mathcal{O}(n_2/r)$ and the low-rank matrix satisfies appropriate incoherence conditions, accurate outlier identification is achievable, with high probability, using a total number of *scalar, linear* measurements of the matrix on the order of $r^2 + k$, times constant and logarithmic factors. Our major contributions here come in the form of extensions of the approach of [15], [16] to settings where the data is corrupted by additive noise, or where the available data are incomplete. In the noisy setting, we describe and analyze a randomized sampling and inference procedure that successfully locates outliers (with high probability) using an effective sampling rate of $\frac{\#_{\text{obs}}}{n_1 n_2} = \mathcal{O}\left(\frac{(r + \log n_2)(n_2/n_L)\mu_V r \log r}{n_1 n_2} + \frac{\log n_2}{n_1}\right)$; in missing-data settings, we present a procedure that succeeds whp using an effective sampling rate $\frac{\#_{\text{obs}}}{pn_1 n_2} = \mathcal{O}\left(\frac{r\mu_L \log^2 n_2}{pn_1}\right)$, where n_L is the number of nonzero columns of \mathbf{L} , p is observation rate in the missing-data setting, and μ_V and μ_L are incoherence parameters (defined in the sequel).

B. Algorithm for Noisy Observations

We first consider model (1) for noisy observations, i.e.,

$$\mathbf{M} = \mathbf{L} + \mathbf{C} + \mathbf{N}, \quad (2)$$

Submitted December 7, 2016. The authors are with the Department of Electrical and Computer Engineering at the University of Minnesota – Twin Cities. Author emails: {lix@1661, jdaupt}@umn.edu. The authors graciously acknowledge support from NSF Award CCF-1217751 and DARPA Young Faculty Award N66001-14-1-4047.

Algorithm 1 Robust Adaptive Compressive Outlier Sensing for Noisy Observations (RACOS-N)

Input: \mathbf{M} , $\gamma \in (0, 1)$, $\lambda, \alpha, \varepsilon_1, \varepsilon_2 > 0$, and $q, m \in [n_1]$
Initialize: $\Phi \in \mathbb{R}^{m \times n_1}$, $\Psi \in \mathbb{R}^{q \times m}$ and $\mathbf{S} = \mathbf{I}_{:, \mathbf{S}}$, where
 $\mathbf{S} = \{j \in [n_2] : S_j \stackrel{iid}{\sim} \text{Bernoulli}(\gamma) = 1\}$ and $p = |\mathbf{S}|$

Step 1

Collect Measurements: $\mathbf{Y}_{(1)} = \Phi \mathbf{M} \mathbf{S}$
Solve OP: $\{\hat{\mathbf{L}}, \hat{\mathbf{C}}\} = \underset{\mathbf{L}, \mathbf{C}}{\text{argmin}} \|\mathbf{L}\|_* + \lambda \|\mathbf{C}\|_{1,2}$
s.t. $\|\mathbf{Y}_{(1)} - \mathbf{L} - \mathbf{C}\|_F \leq \varepsilon_1$
Estimate: $\hat{\mathbf{L}}_{(1)}$ by singular value thresholding operation
on $\hat{\mathbf{L}}$, i.e. $\hat{\mathbf{L}}_{(1)} = \hat{\mathbf{U}} \mathcal{D}_\alpha(\hat{\mathbf{\Sigma}}) \hat{\mathbf{V}}^*$

Step 2

Let: $\hat{\mathcal{L}}_{(1)}$ be the linear subspace spanned by col's of $\hat{\mathbf{L}}_{(1)}$
Set: $\mathbf{P}_{\hat{\mathcal{L}}_{(1)}} \triangleq \mathbf{I} - \mathbf{P}_{\hat{\mathcal{L}}_{(1)}}$
Collect Measurements: $\mathbf{Y}_{(2)} = \Psi \mathbf{P}_{\hat{\mathcal{L}}_{(1)}} (\Phi \mathbf{M})$
Set: $\hat{z}_i = \|(\mathbf{Y}_{(2)})_{:,i}\|_2$, if $\|(\mathbf{Y}_{(2)})_{:,i}\|_2 > \varepsilon_2$
Output: $\hat{\mathcal{L}}_C = \{i : \hat{z}_i \neq 0\}$

where \mathbf{N} is a matrix of additive noise. The key insight in our two-step approach here follows our initial work of [15], and can be described qualitatively as follows.

We consider throughout a column-wise compressed version $\Phi \mathbf{M}$ having many fewer rows than \mathbf{M} , but which still takes the form of a column-wise corrupted low-rank matrix (with the corrupted columns in the same locations as those of the original matrix). In the first step, we apply an existing robust PCA approach designed to be robust to column outliers – called Outlier Pursuit (OP) [9] – to a matrix comprised of a small random subset of columns of $\Phi \mathbf{M}$. This results, in part, in an estimate of the low-rank component $\Phi \mathbf{L}$ of $\Phi \mathbf{M}$, and we identify the subspace spanned from this estimate by the singular vectors of $\Phi \mathbf{L}$ corresponding to singular values above a specified threshold (this serves to mitigate the effects of the noise in the subspace estimate).

Then, a second step incorporates into the sampling operation a composition of an orthogonal projection onto the orthogonal complement of the learned subspace from the first step and an additional column-wise dimensionality reduction operation. This is designed to remove the low-rank component from $\Phi \mathbf{M}$, and to further reduce the dimension of the acquired data. Finally, outlier identification is performed by identifying the columns of the resulting matrix having sufficiently large residual energies. This approach, called Robust Adaptive Compressive Outlier Sensing for noisy observations (RACOS-N), is summarized as Algorithm 1.

C. Algorithm for Incomplete Observations

We also consider variants of the outlier identification problem when the matrix \mathbf{M} has missing elements, where

$$\mathbf{M} = \mathbf{P}_\Omega (\mathbf{L} + \mathbf{C}), \quad (3)$$

and \mathbf{P}_Ω is an operator that masks its arguments that are not in the index set $\Omega \subseteq \{1, 2, \dots, n_1\} \times \{1, 2, \dots, n_2\}$.

Here we employ an analogous approach as in the noisy case. Namely, we operate throughout on a column-wise compressed matrix $\Phi \mathbf{M}$, but consider specifically the case where Φ is a

Algorithm 2 Robust Adaptive Compressive Outlier Sensing for Incomplete Observations (RACOS-I)

Input: \mathbf{M} , Ω , $\gamma_1, \gamma_2 \in (0, 1)$, ρ and $\lambda > 0$
Initialize: $\Phi = \mathbf{I}_{\mathbf{S}_1, :}$ and $\mathbf{S} = \mathbf{I}_{:, \mathbf{S}_2}$, where
 $\mathbf{S}_1 = \{i \in [n_1] : S_i \stackrel{iid}{\sim} \text{Bernoulli}(\gamma_1) = 1\}$, $m = |\mathbf{S}_1|$,
 $\mathbf{S}_2 = \{j \in [n_2] : S_j \stackrel{iid}{\sim} \text{Bernoulli}(\gamma_2) = 1\}$, $\tilde{n}_2 = |\mathbf{S}_2|$.

Step 1

Collect Measurements: $\mathbf{Y}_{(1)} = \Phi \mathbf{M} \mathbf{S}$
Trimming (Optional):
for $j = 1$ **to** \tilde{n}_2
 if # of observed entries of $(\mathbf{Y}_{(1)})_{:,j} > \rho m$
 Select: ρm entries of $(\mathbf{Y}_{(1)})_{:,j}$ uniformly randomly
 Set : The rest entries of $(\mathbf{Y}_{(1)})_{:,j}$ unobserved
 end for
Set: $\Omega_{(1)}$ be the set of observed entries of $\mathbf{Y}_{(1)}$
Solve MP: $\{\hat{\mathbf{L}}_{(1)}, \hat{\mathbf{C}}_{(1)}\} = \underset{\mathbf{L}, \mathbf{C}}{\text{argmin}} \|\mathbf{L}\|_* + \lambda \|\mathbf{C}\|_{1,2}$
s.t. $\mathbf{Y}_{(1)} = \mathbf{P}_{\Omega_{(1)}} (\mathbf{L} + \mathbf{C})$

Step 2

for $j = 1$ **to** n_2 **do**
 Let: $\hat{\mathcal{L}}_{\mathcal{I}_j}$ be subspace spanned by col's of $(\hat{\mathbf{L}}_{(1)})_{\mathcal{I}_j, :}$
 Set: $\mathbf{P}_{\hat{\mathcal{L}}_{\mathcal{I}_j}} \triangleq \mathbf{I} - \mathbf{P}_{\hat{\mathcal{L}}_{\mathcal{I}_j}}$
 Form: $\hat{z}_j = \|\mathbf{P}_{\hat{\mathcal{L}}_{\mathcal{I}_j}} (\Phi \mathbf{M})_{\mathcal{I}_j, j}\|_2$
 end for
Output: $\hat{\mathcal{L}}_C = \{i : \hat{z}_i \neq 0\}$

row submatrix of the $n_1 \times n_1$ identity matrix. In the first step, we apply an existing robust PCA approach designed to be robust to column outliers and missing data – called Manipulator Pursuit (MP) [14], [26] – to a matrix comprised of a small random subset of columns of $\Phi \mathbf{M}$. This results, in part, in an estimate of the low-rank component $\Phi \mathbf{L}$, which we denote by $\hat{\mathbf{L}}_{(1)}$. An optional trimming procedure can be applied before MP by throwing away some entries randomly. This provides better performance for adversarial outliers and improved sampling complexities [14] (see also our additional discussion in Section II).

Then, the second step entails a missing data variant of the orthogonal projection discussed above. Namely, for each $j \in \{1, 2, \dots, n_2\}$, we let $\mathcal{I}_j \subset \{1, 2, \dots, n_1\}$ denote the indices of the observed elements of the j -th column of $\Phi \mathbf{M}$. Then, for each j , we project the observed subvector of the j -th column of $\Phi \mathbf{M}$ onto the orthogonal complement of the column space of a row submatrix of $\hat{\mathbf{L}}_{(1)}$, indexed also by the rows in \mathcal{I}_j . A column is recognized as an outlier if its energy after this orthogonal projection is nonzero. We call the algorithm RACOS for incomplete observation (RACOS-I), and summarize it in Algorithm 2.

D. Comparison with Existing Works

Popular contemporary models for outlier identification based on robust subspace estimation with convex optimization include outlier pursuit (OP) [9], robust computation of linear models (REAPER) [13], and sparse subspace clustering with outliers (SSC) [11]. This is by no means a comprehensive list; see also [27] for a survey of more classical methods in robust statistics. Here we focus on comparing these contemporary

TABLE I
COMPARISON OF DIFFERENT OUTLIER IDENTIFICATION MODELS IN TERMS OF ASSUMPTIONS ON \mathbf{L} AND \mathbf{C} , COMPUTATIONAL FORMULATION, EXISTENCE OF GUARANTEES FOR MODELING WITH MISSING ENTRIES, AND WHETHER THE GUARANTEES ARE PROBABILISTIC IN NATURE

Model	OP	REAPER	SSC
Assump. on \mathbf{L}	Incoherence	Permeance	Incoherence
Assump. on \mathbf{C}	$k = \mathcal{O}(1/r)$	Small Alignment	Isotropy
Computation	Convex	Convex (Relax.)	Convex
Missing Entry	Yes [14], [33]	Unknown	Unknown
Prob. Result	No	Both	Yes

models in terms of the model assumption, recovery performance, computational efficiency, the tolerance of missing entries, etc.

OP assumes low coherence of inliers and the number of outliers to be small. While, it does not require outliers to be isotropic and also has guarantees for observation with missing entries (Manipulator Pursuit (MP), [14]). REAPER forms a convex relaxation of least orthogonal absolute deviations [28] and defines several summary statistics explicitly, including permeance (large evidence of inliers), total inlier residual (small deviation/noise on inliers), and alignment (small collinearity of outliers), to reveal the effectiveness of the model. The analysis of REAPER include both the deterministic model and a (Haystack) random model. SSC also assumes incoherence of inliers and isotropy of outliers, where the incoherence here is slightly different with the incoherence in OP, as SSC considers union of subspaces. Though SSC has an algorithmic extension for incomplete observations [29], neither REAPER nor SSC has theoretical guarantees for incomplete observations with outliers so far. In addition, SSC provides a probabilistic result as it considers random outliers [11].

A summary of the properties of these models are provided in Table I. Due to the relatively intuitive condition on \mathbf{L} and our interest in the case that only a small number of outliers exist (no alignment or uniformity of \mathbf{C} is required), we use OP as our underlying model that can also handle the case of incomplete observations. Subsequent to our initial work, a downsampling-based approach of Robust PCA was proposed in [30] for noiseless and complete observation, where the goal includes recovery of the subspace spanned by the columns of the low-rank component. Alternative approaches to the robust subspace recovery problem, e.g., [11], [13], [31], [32], can also be incorporated into our overall approach instead of OP, and could yield potential improvements (e.g., in terms of the structural assumptions under which recovery is guaranteed). Investigations along these lines are left for future effort.

E. Notation

Bold-face upper-case letters (\mathbf{M} , Φ etc.) are used to denote matrices, bold-face lower-case letters (\mathbf{x} , \mathbf{v} , etc.) to denote vectors, and non-bold letters are used to denote scalar parameters or constants. We employ both ‘block’ and ‘math’ type notations (e.g., \mathbf{L}, \mathbf{L}), where the latter are used to denote variables in the optimization tasks. Given a positive integer n , we denote $[n] \triangleq \{1, 2, \dots, n\}$.

The ℓ_p norm of a vector $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]$ is $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. For a matrix \mathbf{X} , we denote the nuclear norm

(sum of singular values) by $\|\mathbf{X}\|_*$, the spectral norm (largest singular value) by $\|\mathbf{X}\|_2$, the $\ell_{1,2}$ norm (sum of column ℓ_2 norms) by $\|\mathbf{X}\|_{1,2}$, and the $\ell_{\infty,2}$ norm (largest column ℓ_2 norm) by $\|\mathbf{X}\|_{\infty,2}$.

For a rank r matrix \mathbf{L} , we denote the compact singular value decomposition (SVD) of \mathbf{L} as $\mathbf{L} = \mathbf{U}\Sigma\mathbf{V}^*$, where $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times r}$ have orthonormal columns, and $\Sigma \in \mathbb{R}^{r \times r}$ is diagonal with the i -th largest singular value $\sigma_i(\mathbf{L})$ of \mathbf{L} as the i -th diagonal element, i.e. $\Sigma_{ii} = \sigma_i(\mathbf{L})$. We also denote $\mathbf{P}_{\mathcal{L}}(\mathbf{X}) = \mathbf{U}\mathbf{U}^T\mathbf{X}$ and $\mathbf{P}_{\mathcal{L}^\perp}(\mathbf{X}) = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{X}$ as projection operations that project a matrix \mathbf{X} onto the column space and the orthogonal complement of column space of \mathbf{L} respectively. The mask operator $\mathbf{P}_\Omega(\cdot)$ is defined via $(\mathbf{P}_\Omega(\mathbf{X}))_{ij} = \mathbf{X}_{ij}$, if $(i, j) \in \Omega$; or 0 if $(i, j) \notin \Omega$ given a support set $\Omega \subseteq [n_1] \times [n_2]$.

MATLAB-inspired notation is used to denote submatrices; e.g., $\mathbf{I}_{\mathcal{S},:}$ (or $\mathbf{I}_{:, \mathcal{S}}$) is used to denote the submatrix formed by extracting rows (or columns) of \mathbf{I} indexed by \mathcal{S} . Likewise, we use $\mathbf{X}_{:,j}$ to denote j -th column of \mathbf{X} .

II. MAIN RESULTS

Here we provide the theoretical guarantees of RACOS for model (2) and (3). The noisy observation setting with a generic additive noise will be discussed first. Then, we specialize this to a setting where the noise is random. Finally, we provide the result for the incomplete observation setting with and without a ‘trimming’ step.

A. Preliminary Assumptions

We first introduce two important properties on which our recovery guarantees are based. It is well-known that the decomposition of a matrix into a low-rank component and a sparse component is a ill-posed problem in general. For example, a matrix with only one non-zero entry is both a low-rank and sparse matrix. The first property is a widely adopted notion of ‘incoherence’ in the literature of robust PCA [7]–[9] to overcome such identifiability issues.

Definition II.1 (Row and Column Incoherence Properties). *Let $\mathbf{L} \in \mathbb{R}^{n_1 \times n_2}$ be a rank r matrix with at most $n_{\mathbf{L}} \leq n_2$ nonzero columns. Given the compact SVD $\mathbf{L} = \mathbf{U}\Sigma\mathbf{V}^*$, \mathbf{L} is said to satisfy the **row incoherence property** with parameter $\mu_{\mathbf{U}} \in [1, n_1/r]$ if*

$$\max_{i \in [n_1]} \|\mathbf{U}^* \mathbf{e}_i\|_2^2 \leq \mu_{\mathbf{U}} \frac{r}{n_1},$$

where $\{\mathbf{e}_i\}$ are canonical basis vectors for \mathbb{R}^{n_1} . Likewise, \mathbf{L} is said to satisfy the **column incoherence property** with parameter $\mu_{\mathbf{V}} \in [1, n_{\mathbf{L}}/r]$ if

$$\max_{j \in [n_2]} \|\mathbf{V}^* \mathbf{e}_j\|_2^2 \leq \mu_{\mathbf{V}} \frac{r}{n_{\mathbf{L}}},$$

where $\{\mathbf{e}_j\}$ are canonical basis vectors for \mathbb{R}^{n_2} .

The second important property is a criteria for the random measurement matrices to preserve the Euclidean norm of any fixed vector with high probability, which we formalize as follows.

Definition II.2 (Distributional Johnson-Lindenstrauss (JL) Property). A (random) matrix $\Phi \in \mathbb{R}^{m \times n}$ is said to satisfy the **distributional JL property** if for any fixed $\mathbf{v} \in \mathbb{R}^n$ and any $\varepsilon \in (0, 1)$,

$$\Pr \left(\left| \|\Phi \mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2 \right| \geq \varepsilon \|\mathbf{v}\|_2^2 \right) \leq 2e^{-mf(\varepsilon)}, \quad (4)$$

where $f(\varepsilon) > 0$ is a constant depending only on ε that is specific to the distribution of Φ .

B. Guarantees for Noisy Observations

1) *Structural Assumptions*: Motivated from our work of the noiseless case in [15], we state the *structural conditions* for noisy observations as following:

- (d1) $\text{rank}(\mathbf{L}) = r < \min\{n_1, n_2\}$,
- (d2) \mathbf{L} has $n_L = n_2 - k$ nonzero columns,
- (d3) \mathbf{L} satisfies the *column incoherence property* with parameter $\mu_{\mathbf{V}}$,
- (d4) the condition number of \mathbf{L} satisfies $\kappa = \frac{\sigma_1(\mathbf{L})}{\sigma_r(\mathbf{L})} < \infty$, and
- (d5) \mathbf{C} has $|\mathcal{I}_{\mathbf{C}}| = k$ nonzero columns, where $\mathcal{I}_{\mathbf{C}} \triangleq \{i \in [n_2] : \|\mathbf{P}_{\mathcal{L}^\perp} \mathbf{C}_{:,i}\|_2 > \tau_1 \|\mathbf{C}_{:,i}\|_2\}$ for some constant $\tau_1 \in (0, 1)$.

The conditions (d1)-(d3) are natural for \mathbf{L} , and are similar to those imposed in our prior work [15]. Note that $n_L + k \leq n_2$ in general, though we restrict our attention here to the case $n_L + k = n_2$. Without loss of generality (w.l.o.g.), we assume that for the inlier columns of \mathbf{L} , the corresponding columns of \mathbf{C} are 0, and for outlier columns of \mathbf{C} , the corresponding columns of \mathbf{L} are zero. The condition (d4) assumes the well-conditioning of the low-rank matrix \mathbf{L} , which is a mild assumption in practice, and (d5) is a condition on the outlier columns that can be viewed as a slightly stronger version of our analogous property assumed in [15] for the noise-free case. That the residuals of the outlier columns need to be sufficiently large is somewhat intuitive, as the outlier columns projected onto the complement of the low-rank subspace need to be large enough due to the inexact estimate of noisy low-rank subspace. In our analysis, the quantity τ_1 is proportional to the upper bound of the estimation error (in spectral norm) of the low-rank subspace, which is simply 0 under analogous structural assumptions when $\mathbf{N} = \mathbf{0}_{n_1 \times n_2}$. Therefore (d5) represents a natural extension of the analogous condition imposed in [15] for the noise-free case.

We also impose conditions on the noise to facilitate exact outlier detection. Indeed if, for example, \mathbf{N} has very large Euclidean norm in some column, then we may confuse it for a true outlier column. To avoid such undesirable situations, we impose several conditions on the noise, and its relationship to \mathbf{L} and \mathbf{C} . For notational simplicity, we define

$$\eta_{\mathbf{N}} = \max_{j \in [n_2]} \|\mathbf{N}_{:,j}\|_2.$$

Then the structural conditions of \mathbf{N} are as following:

- (n1) $\sigma_r(\mathbf{L}) > \frac{90\sqrt{2}\gamma}{\tau_1} n_2 \eta_{\mathbf{N}}$, and
 - (n2) $\min_{i \in \mathcal{I}_{\mathbf{C}}} \|\mathbf{C}_{:,i}\|_2 > \tau_2 \eta_{\mathbf{N}}$ for some constant τ_2 ,
- where $\gamma \in (0, 1)$ is the column subsampling parameter (an input to Algorithm 1). The condition (n1) is akin to an SNR assumption, and ensures that all singular values of \mathbf{L} dominate

the Euclidean norm of columns of \mathbf{N} . Note that the condition (n1) may seem somewhat restrictive, but both our analysis and numerical evaluation indicate that $\sigma_r(\mathbf{L}) = \Omega(\sqrt{\gamma} n_2 \eta_{\mathbf{N}})$ may be necessary for our approach. The condition (n2) is a structural one that any outlier column dominates the per-column noise. It is interesting to notice that the conditions (n1) and (n2) hold somewhat trivially when $\mathbf{N} = \mathbf{0}_{n_1 \times n_2}$.

2) *Generic Recovery Guarantees*: We first provide a description of the singular value hard value thresholding operation. Specifically, let the SVD of $\hat{\mathbf{L}}$ be $\hat{\mathbf{L}} = \hat{\mathbf{U}} \hat{\Sigma} \hat{\mathbf{V}}^*$ with $\hat{\Sigma} = \text{diag}(\{\hat{\sigma}_i\}_{1 \leq i \leq \min\{m, \tilde{n}_2\}})$, where \tilde{n}_2 is the number of columns of \mathbf{S} . By choosing a constant α , we then apply a singular value thresholding operation defined as $\mathcal{D}_\alpha(\hat{\Sigma}) = \text{diag}(\{f(\hat{\sigma}_i, \alpha)\}_{1 \leq i \leq \min\{m, \tilde{n}_2\}})$, where $f(\cdot, \cdot)$ is

$$f(\hat{\sigma}_i, \alpha) \triangleq \begin{cases} \hat{\sigma}_i, & \text{if } \hat{\sigma}_i > \alpha \\ 0, & \text{if } \hat{\sigma}_i \leq \alpha \end{cases},$$

and the estimate of the low-rank matrix is $\hat{\mathbf{L}}_{(1)} = \hat{\mathbf{U}} \mathcal{D}_\alpha(\hat{\Sigma}) \hat{\mathbf{V}}^*$. In the next theorem, we state our main results for outlier identification for observation under the general additive noise model (2).

Theorem II.1 (Accurate Recovery via RACOS-N). Suppose $\mathbf{M} = \mathbf{L} + \mathbf{C} + \mathbf{N}$, where \mathbf{L} and \mathbf{C} satisfy the structural conditions (d1)-(d5) with the number of outliers k upper bounded by k_u ,

$$k \leq k_u = \frac{1}{3(1 + 1024 r \mu_{\mathbf{V}})} n_2. \quad (5)$$

Let the measurement matrices Φ and Ψ be drawn from any distribution following (4), and for a fixed $\delta \in (0, 1)$, suppose that the column subsampling parameter γ , and the row and column sampling parameters m and q , respectively, satisfy

$$\gamma \geq \max \left\{ \frac{200 \log(\frac{6}{\delta})}{n_L}, \frac{600(1 + 1024 r \mu_{\mathbf{V}}) \log(\frac{6}{\delta})}{n_2}, \frac{10 r \mu_{\mathbf{V}} \log(\frac{6r}{\delta})}{n_L} \right\}, \quad (6)$$

$$m \geq \frac{5(r+1) + \log(2n_2) + \log \frac{2}{\delta}}{f(1/4)}, \quad (7)$$

$$q \geq \frac{4 \log \frac{2n_2}{\delta}}{f(1/4)}. \quad (8)$$

Further suppose that \mathbf{N} satisfies the structural conditions (n1) and (n2), where the constant τ_2 satisfies

$$\tau_1 \tau_2 > 6(\beta + 1)(\tau_1/4 + 1) + 90\sqrt{6}\gamma\beta\kappa n_2, \quad (9)$$

with a constant $\beta > \sqrt{3}$, and the regularization parameter λ in OP satisfies

$$\lambda = \frac{3\sqrt{1 + 1024\mu_{\mathbf{V}}r}}{14\sqrt{\tilde{n}_2}}, \quad (10)$$

where \tilde{n}_2 is the number of columns of \mathbf{S} . Then there exist a singular value hard thresholding constant α and an ε_2 satisfying

$$18\gamma n_2 \eta_{\mathbf{N}} < \alpha < 54\gamma n_2 \eta_{\mathbf{N}}, \quad (11)$$

$$\max_{j \in \mathcal{I}_{\mathbf{L}}} \|\Psi \mathbf{P}_{\hat{\mathcal{L}}_{(1)}^\perp}(\Phi \mathbf{M}_{:,j})\|_2 < \varepsilon_2 < \min_{i \in \mathcal{I}_{\mathbf{C}}} \|\Psi \mathbf{P}_{\hat{\mathcal{L}}_{(1)}^\perp}(\Phi \mathbf{M}_{:,i})\|_2, \quad (12)$$

such that the following claims hold simultaneously with probability at least $1 - 3\delta$:

- (C1) RACOS-N correctly identifies the salient columns of \mathbf{C} (i.e., $\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}}$), and
 (C2) the total number of measurements collected is no greater than $((\frac{3}{2})\gamma m + q)n_2$.

It is interesting to note that the sufficient condition (5) on the number of identifiable outliers is of the same order compared with OP [9] and noiseless ACOS [15], which can be as large as a fixed proportion of n_2 when both the rank r and column coherence parameter $\mu_{\mathbf{V}}$ are small. In terms of the sample complexity, we show that our approach succeeds with high probability with effective sampling rate $\frac{\#_{\text{obs}}}{n_1 n_2} = \mathcal{O}\left(\frac{(r+\log n_2)(n_2/n_{\mathbf{L}})\mu_{\mathbf{V}}r \log r}{n_1 n_2} + \frac{\log n_2}{n_1}\right)$. This may potentially be much smaller than 1 when, e.g., r is small relative to the problem dimensions.

We also present the performance guarantees for RACOS-N when we simply take the column-wise Euclidean norms in Step2, i.e. $\Psi = \mathbf{I}$ is an identity matrix, in the following corollary. The analysis follows directly from that of Theorem II.1, thus we omit it here.

Corollary II.1. Suppose all conditions in Theorem II.1 hold, except that Ψ is an identity matrix, i.e. $q = m$, and the constant τ_2 satisfies (9) with a constant $\beta > 1$. If λ satisfies (10), then for α and ε_2 satisfying (11) and (12) respectively, the following claims hold simultaneously with probability at least $1 - 2\delta$:

- (C3) RACOS-N correctly identifies the salient columns of \mathbf{C} (i.e., $\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}}$), and
 (C4) the total number of measurements collected is no greater than mn_2 .

We can see that the recoverability of RACOS-N in terms of the noise for $\Psi = \mathbf{I}$ is stronger than that when Ψ is a random matrix, where we require a smaller lower bound for τ_2 , hence smaller lower bound requirement for $\min_{i \in \mathcal{I}_{\mathbf{C}}} \|\mathbf{C}_{:,i}\|_2$ when $\Psi = \mathbf{I}$. This is intuitively reasonable since fewer random projections facilitate less ambiguity of the original data. On the other hand, the overall sample complexity for random Ψ is $\mathcal{O}(\gamma mn_2 + qn_2)$, which is potentially much smaller than $\mathcal{O}(mn_2)$ for $\Psi = \mathbf{I}$, when γ and q are small. This can be viewed as a trade off between the outlier detection performance and the sample complexity. Further improvement of sample complexity can be achieved using multivariate regression [34] and the grouping idea [35], if the grouping structure exists among the outliers.

3) *Observations with Random Noise:* We now consider the observation setting (2) with a random noise \mathbf{N} . Specifically, we assume that \mathbf{N} has i.i.d. zero-mean Gaussian entries, in which case we can specify the singular value hard thresholding constant α . The following Theorem quantifies the constant for the observation with a Gaussian noise. The proof is provided in Appendix VI-D. The analysis can be extended to other type of random noises, such as subgaussian entries, in a straightforward manner.

Theorem II.2. Suppose $\mathbf{M} = \mathbf{L} + \mathbf{C} + \mathbf{N}$, where \mathbf{L} and \mathbf{C} satisfy the structural conditions (d1)-(d5) with k satisfying (5). Also suppose for any $\delta \in (0, 1)$, (6) holds, and the

measurement matrices Φ and Ψ are drawn from a distribution satisfying (4), (7) and (8). Further suppose \mathbf{N} has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries and satisfies conditions (n1) and (n2) with the constant τ_2 satisfying (9). If the regularization parameter λ satisfies (10), then for α satisfying

$$18C_1\gamma\sigma n_2 < \alpha < 54C_2\gamma\sigma n_2, \quad (13)$$

where $C_1 = \left(n_1 - (8n_1 \log \frac{2n_2}{\delta})^{1/2}\right)^{1/2}$ and $C_2 = \left(n_1 + (8n_1 \log \frac{2n_2}{\delta})^{1/2}\right)^{1/2}$, and ε_2 satisfying (12), claims (C1) and (C2) hold simultaneously with probability at least $1 - 4\delta$.

For completeness, we also present the performance guarantees for RACOS-N with $\Psi = \mathbf{I}$ under the Gaussian noise \mathbf{N} in the following corollary without proof.

Corollary II.2. Suppose all conditions in Theorem II.2 hold, except that Ψ is an identity matrix, i.e. $q = m$, and the constant τ_2 satisfies (9). If λ satisfies (10), then for α and ε_2 satisfying (13) and (12) respectively, claims (C3) and (C4) hold simultaneously with probability at least $1 - 3\delta$.

We investigate the implications of these results experimentally in Section IV.

C. Guarantees for Incomplete Observations

We now consider the “missing data” setting.

1) *Structural Assumptions:* For notational simplicity, we denote $\mathbf{C}_{\Omega} = \mathbf{P}_{\Omega}(\mathbf{C})$. We state the structural conditions for the incomplete observation setting (3) as following (adapted from [15]):

- (g1) $\text{rank}(\mathbf{L}) = r$,
 (g2) \mathbf{L} has $n_{\mathbf{L}}$ nonzero columns,
 (g3) \mathbf{L} satisfies the row and column incoherence properties with parameters $\mu_{\mathbf{U}}$ and $\mu_{\mathbf{V}}$ respectively, and
 (g4) \mathbf{C} has $|\mathcal{I}_{\mathbf{C}}| = k$ nonzero columns, where $\mathcal{I}_{\mathbf{C}} \triangleq \{j \in [n_2] : \forall \mathcal{I}^* \subset [n_1] \text{ with } |\mathcal{I}^*| \geq \frac{r\mu_{\mathbf{U}} \log(2r)}{p}, \|\mathbf{P}_{\mathcal{I}^*}^{\perp}(\mathbf{C}_{\Omega})\|_{\mathcal{I}^*,j}\|_2 > 0\}$.

Note that the low-rank matrix \mathbf{L} need to satisfy both column and row incoherence properties due to simultaneous column and row sampling procedure. The condition (g4) is from the fact that we only need to consider those observed entries in outlier columns. For missing entries in outlier columns, we will never be able to recover them exactly.

2) *Generic Recovery Guarantees:* In the following theorem, we state our main result for model (3) using RACOS-I without trimming.

Theorem II.3 (Accurate Recovery via RACOS-I without Trimming). Suppose $\mathbf{M} = \mathbf{P}_{\Omega}(\mathbf{L} + \mathbf{C})$, where the components \mathbf{L} and \mathbf{C} satisfy the structural conditions (g1)-(g4). Let $\mu_{\mathbf{L}} = \max(\mu_{\mathbf{U}}, \mu_{\mathbf{V}})$. Assume p and k satisfy

$$p \geq p_l = \frac{C_p \mu_{\mathbf{L}}^2 r^2 \log^3(4n_{\mathbf{L}})}{n_1}, \quad (14)$$

$$k \leq k_u = \frac{p^2 n_2 / 3}{p^2 + C_k (1 + \frac{3\sqrt{6}\mu_{\mathbf{L}}r}{p\sqrt{n_1}}) \mu_{\mathbf{L}}^3 r^3 \log^6(4n_{\mathbf{L}})}, \quad (15)$$

for some positive constants C_p and C_k . Given any $\delta \in (0, 1/2)$, if $n_L \geq \frac{\delta e^{8p}}{4}$, the row sampling parameter γ_1 and column sampling parameter γ_2 satisfy

$$\gamma_1 \geq \max \left\{ \frac{2r\mu_U \log(2r)}{n_1 p}, \frac{8 \log \frac{4n_L}{\delta}}{n_1 p}, \frac{10r\mu_U \log \frac{4r}{\delta}}{n_1}, \frac{162p_l}{p} \right\}, \quad (16)$$

$$\gamma_2 \geq \max \left\{ \frac{200 \log(\frac{9}{\delta})}{n_L}, \frac{10r\mu_V \log(\frac{9r}{\delta})}{n_L}, \frac{C_{\gamma_2}(\frac{1}{\delta})^{\frac{1}{5}}}{n_2}, \frac{200 \log(\frac{9}{\delta})}{k_u} \right\}, \quad (17)$$

for some positive constant C_{γ_2} , and the regularization parameter in MP satisfies

$$\lambda = \frac{1}{48} \sqrt{\frac{p}{9kr\mu_L \log^2(4\gamma_2 n_L)}}, \quad (18)$$

then the following claims hold simultaneously with probability at least $1 - 2\delta$:

- (C5) RACOS-I correctly identifies the salient columns of \mathbf{C} (i.e., $\tilde{\mathcal{I}}_C = \mathcal{I}_C$), and
- (C6) the total number of measurements collected is no greater than $\frac{3}{2}p\gamma_1 n_1 n_2$.

We see that the sampling complexity reduces from $pn_1 n_2$ for the full model to $O(p\gamma_1 n_1 n_2)$ for RACOS-I, which is significant if $r \ll \max\{n_1, n_2\}$. In terms of computational complexity, RACOS-I reduces the size of the matrix operated in MP from $n_1 n_2$ to $\gamma_1 \gamma_2 n_1 n_2$, and the computational cost in each iteration of MP (in the proximal first order algorithm), dominated by SVD, reduces from $\mathcal{O}(n_1 n_2 \min\{n_1, n_2\})$ to $\mathcal{O}(\gamma_1 \gamma_2 n_1 n_2 \min\{\gamma_1 n_1, \gamma_2 n_2\})$. This improvement is significant if r is small. Note that the last terms in both (16) and (17) are the dominating terms, which can be improved by the trimming procedure. In the next theorem, we provide the main result for of RACOS-I with trimming.

Theorem II.4 (Accurate Recovery via RACOS-I with Trimming). *Let $\varphi = \frac{p}{p}$. Suppose $\mathbf{M} = \mathbf{P}_\Omega(\mathbf{L} + \mathbf{C})$, where the components \mathbf{L} and \mathbf{C} satisfy the structural conditions (g1)-(g4). Let $\mu_L = \max(\mu_U, \mu_V)$. Assume p and k satisfy*

$$p \geq p_l = C_p \left(1 + \frac{1}{\varphi}\right) \frac{\mu_L r \log^2(2n_2)}{n_1}, \quad (19)$$

$$k \leq k_u = C_k \frac{\varphi}{1 + \varphi\sqrt{\varphi}} \frac{pn_L}{\mu_L^{3/2} r^{3/2} \log^3(2n_2)}, \quad (20)$$

for some positive constants C_p and C_k . Given any $\delta \in (0, 1/2)$, if the row sampling parameter γ_1 and the column sampling parameter γ_2 satisfy (16) and (17) respectively, and the regularization parameter satisfies

$$\lambda = \frac{1}{48} \sqrt{\frac{1}{\sqrt{(1 + \varphi)r\mu_L k \log(n_1 + n_L)}}}, \quad (21)$$

then claims (C5) and (C6) hold simultaneously with probability at least $1 - 2\delta$.

From Theorem II.3, RACOS-I without trimming reduces the dimension of the matrix operated in MP from $n_1 n_2$ to $\mathcal{O}(\mu_L^5 r^5 \log^9 n_2 / p^3)$. On the other hand, from Theorem II.4, RACOS-I with trimming reduces the dimension of the matrix

operated in MP from $n_1 n_2$ to $\mathcal{O}(\mu_L^{5/2} r^{5/2} \log^5 n_2 / p^2)$. It is also demonstrated in [14] that (19) and (20) are close to information-theoretic (minimax) optimal, where the trimming is step is crucial in the analysis. We refer interested reader to [14] for further discussion. Though, RACOS-I with trimming has stronger theoretical guarantees, it has one more parameter ρ to choose. Thus, in practice, we take the trimming as an option to trade off the ease of the algorithmic procedure and performance of sampling complexity. As with the noisy case, we evaluate the implications of these results experimentally in Section IV.

III. PROOF OF MAIN RESULTS

In this section, we provide the sketch of the proof for Theorem II.1, which is formalized by the following intermediate lemmata. The proofs of the lemmata are deferred to the appendix. The proofs for Theorem II.3 and Theorem II.4 are analogous to the proof for Theorem II.1, and are deferred to the supplemental material.

For notional convenience, we first introduce:

$$\tilde{\mathbf{M}} \triangleq \Phi \mathbf{M} = \Phi \mathbf{L} + \Phi \mathbf{C} + \Phi \mathbf{N} = \tilde{\mathbf{L}} + \tilde{\mathbf{C}} + \tilde{\mathbf{N}}, \quad (22)$$

We begin by validating that if conditions (d1)-(d5) and (n1)-(n2) hold, then analogous structural conditions also hold for $\tilde{\mathbf{M}}$ provided that m is sufficiently large. This is stated as Lemma III.1, and we provide the proof in Appendix VI-A.

Lemma III.1. *Suppose $\mathbf{M} = \mathbf{L} + \mathbf{C} + \mathbf{N}$, where \mathbf{L} and \mathbf{C} satisfy the structural conditions (d1)-(d5), and \mathbf{N} satisfies conditions (n1) and (n2). Given $\delta \in (0, 1)$, further suppose Φ is an $m \times n_1$ matrix drawn from a distribution satisfying the distributional JL property (4) with m satisfying (7), and let $\tilde{\mathbf{M}} = \tilde{\mathbf{L}} + \tilde{\mathbf{C}} + \tilde{\mathbf{N}}$ be as defined in (22). Then, with probability at least $1 - \delta$, the components $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{C}}$ satisfy*

- (d1) $\text{rank}(\tilde{\mathbf{L}}) = r$,
 - (d2) $\tilde{\mathbf{L}}$ has n_L nonzero columns,
 - (d3) $\tilde{\mathbf{L}}$ satisfies the column incoherence property with parameter μ_V ,
 - (d4) condition number of $\tilde{\mathbf{L}}$ satisfies $\frac{\sigma_1(\tilde{\mathbf{L}})}{\sigma_r(\tilde{\mathbf{L}})} \leq \sqrt{3}\kappa$, and
 - (d5) $\tilde{\mathbf{C}}$ has $|\mathcal{I}_{\tilde{\mathbf{C}}}| = k$ nonzero columns, where $\mathcal{I}_{\tilde{\mathbf{C}}} \triangleq \{i \in [n_2] : \|\mathbf{P}_{\tilde{\mathbf{L}}^\perp} \tilde{\mathbf{C}}_{:,i}\|_2 > \tau_1 \|\tilde{\mathbf{C}}_{:,i}\|_2 / 2\}$ and $\mathcal{I}_{\tilde{\mathbf{C}}} = \mathcal{I}_C$.
- Simultaneously, let $\eta_{\tilde{\mathbf{N}}} = \max_{j \in [n_2]} \|\tilde{\mathbf{N}}_{:,j}\|_2$, which satisfies

$$\frac{4}{5} \eta_{\tilde{\mathbf{N}}} \leq \eta_{\tilde{\mathbf{N}}} \leq \frac{6}{5} \eta_{\tilde{\mathbf{N}}}, \quad (23)$$

then we further have

- (n1) $\sigma_r(\tilde{\mathbf{L}}) > \frac{\tau_2 \sqrt{2\gamma}}{\tau_1} n_2 \eta_{\tilde{\mathbf{N}}}$, and
- (n2) $\min_{i \in \mathcal{I}_C} \|\tilde{\mathbf{C}}_{:,i}\|_2 > \frac{4}{5} \tau_2 \eta_{\tilde{\mathbf{N}}}$.

Now suppose conditions (d1)-(d5), (n1) and (n2) hold. We then establish that the number of columns generated by the column downsampling matrix \mathbf{S} is close to γn_2 , and Step 1 of Algorithm 1 approximately preserves the column space $\tilde{\mathcal{L}}$ of $\tilde{\mathbf{L}}$ such that the contaminated outlier columns have larger residuals than that of the inlier columns after the orthogonal projection onto the complement of the estimated low-rank

subspace. This is formalized in Lemma III.2 and we provide the proof in Appendix VI-B.

Lemma III.2. Let $\tilde{\mathbf{M}} = \tilde{\mathbf{L}} + \tilde{\mathbf{C}} + \tilde{\mathbf{N}}$ be an $m \times n_2$ matrix, where the components $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{C}}$ satisfy the conditions (d1)-(d5). For any $\delta \in (0, 1)$, suppose the column sampling parameter γ satisfies (6), and $\tilde{\mathbf{N}}$ satisfies (n1) and (n2) with k satisfying (5). Further suppose τ_2 satisfies (9). If λ satisfies (10), and the singular value hard thresholding constant α satisfies (11), then the following claims hold simultaneously with probability at least $1 - \delta$:

- (I) \mathbf{S} has $\tilde{n}_2 \leq (3/2)\gamma n_2$ columns, and
- (II) for any $i \in \mathcal{I}_{\mathbf{C}}$ and $j \in \mathcal{I}_{\mathbf{L}}$, we have that

$$\|\mathbf{P}_{\hat{\mathcal{L}}_{(1)}}^\perp(\tilde{\mathbf{C}} + \tilde{\mathbf{N}})_{:,i}\|_2 > \beta \|\mathbf{P}_{\hat{\mathcal{L}}_{(1)}}^\perp(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2. \quad (24)$$

The last intermediate result is to show that Step 2 of Algorithm 1 produces the correct set of outlier columns with high probability, provided that (24) holds. This is summarized in Lemma III.3 and its proof is provided in Appendix VI-C.

Lemma III.3. For any $\delta \in (0, 1)$, suppose (24) holds, and $\Psi \in \mathbb{R}^{q \times m}$ is a matrix drawn from a distribution satisfying the distributional JL property (4) with q satisfying (8). If ε_2 satisfies (12), then with probability at least $1 - \delta$, we have $\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}}$ from Step 2 of Algorithm 1.

The overall results of Theorem II.1 follows by combining three intermediate results provided in Lemma III.1, Lemma III.2, and Lemma III.3, using the union bound. Therefore, with probability at least $1 - 3\delta$, the claims (C1) and (C2) of Theorem II.1 hold.

IV. EXPERIMENTAL EVALUATION

In this section, we demonstrate explicitly via the numerical evaluation that the sample complexities we derived in the main results are tight in practice¹. We also examine the computational performance to quantify the improvement of our proposed method over the full data models OP and MP. The timing is recorded as the CPU execution time of the algorithm for different combinations of parameters $(m, \gamma, \gamma_1, \gamma_2)$. All results are evaluated by averaging 100 random trials with MATLAB R2014b on an iMac with a 3.4 GHz Intel Core i7 processor, 32 GB memory, and running OS X 10.8.5.

A. Evaluation for Noisy Observation Settings

A trial is deemed a success if the following holds:

$$\min_{i \in \mathcal{I}_{\mathbf{C}}} \|\Psi \mathbf{P}_{\hat{\mathcal{L}}_{(1)}}^\perp(\Phi \mathbf{M}_{:,i})\|_2 > \max_{i \in \mathcal{I}_{\mathbf{L}}} \|\Psi \mathbf{P}_{\hat{\mathcal{L}}_{(1)}}^\perp(\Phi \mathbf{M}_{:,i})\|_2,$$

which implies that there exists a constant threshold ε_2 such that the column-wise hard thresholding yields accurate support recovery. For the singular value hard thresholding, we choose the constant α that preserves 99% of the sum of singular values, which performs well in our settings of evaluations.

¹Outlier recovery transition plots for ACOS are provided in Li & Haupt [15] to demonstrate the recoverability in term of r and k for different levels of noise $\sigma_{\mathbf{N}}$ for noisy observations and different sampling parameter p for observation with missing entries. More results on real data evaluations are provided in [35], [36].

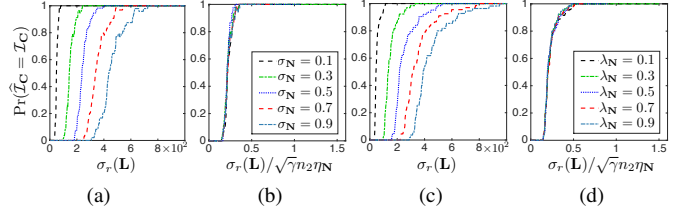


Fig. 1. Demonstration of the probability of success versus the minimal singular value $\sigma_r(\mathbf{L})$ of \mathbf{L} for Gaussian noise under different choices of the variance $\sigma_{\mathbf{N}}$ (a and b) and Laplace noise under different choices of the parameters $\lambda_{\mathbf{N}}$ (c and d). (b) and (d) provide the results with rescaling of $\sigma_r(\mathbf{L})$ by $\sqrt{\gamma n_2 \eta_{\mathbf{N}}}$.

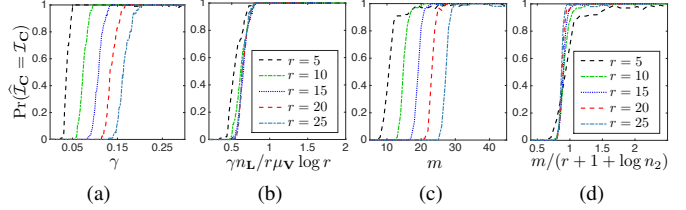


Fig. 2. Demonstration of the probability of success versus column subsample parameter γ (a and b) and row sampling parameter m (c and d) for noisy observations under different settings of rank r of \mathbf{L} . (b) and (d) provide the results with rescaling of γ by $\frac{r \mu \mathbf{V} \log(r)}{n_{\mathbf{L}}}$ and m by $r + 1 + \log k$ respectively.

We generate both the row sampling matrix Φ and the row reduction matrix Ψ with i.i.d. $\mathcal{N}(0, 1)$ entries, without the evaluation of $\Psi = \mathbf{I}$. We fix $n_1 = 100$, $n_2 = 1000$, $q = 20$, $k = 0.2n_2$, $n_{\mathbf{L}} = n_2 - k$, and $\lambda = 0.4$, and justify the claimed bounds via varying parameters, such as r , m and γ .

We first demonstrate that $\sigma_r(\mathbf{L}) = \Omega(\sqrt{\gamma n_2 \eta_{\mathbf{N}}})$ in (n1) appears to be a necessary bound in practice. Let $r = 5$, $m = 0.3n_1$ and $\gamma = 0.2$. We generate two random matrices $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{n_{\mathbf{L}} \times r}$ with i.i.d. $\mathcal{N}(0, 1)$ entries, and take $\mathbf{L}_0 = [\mathbf{U}\mathbf{V}^T \mathbf{0}_{n_1 \times k}]$. Then let $\mathbf{L} = \frac{\sigma_r(\mathbf{L})}{\sigma_r(\mathbf{L}_0)} \mathbf{U}_0 \Sigma_0 \mathbf{V}_0^T$, where $\mathbf{U}_0 \Sigma_0 \mathbf{V}_0^T$ is SVD of \mathbf{L}_0 , $\sigma_r(\mathbf{L}_0) = (\Sigma_0)_{rr}$ is the minimal singular value of \mathbf{L}_0 , and $\sigma_r(\mathbf{L})$ is a parameter to control the singular values of \mathbf{L} . The outlier matrix is generated as $\mathbf{C} = [\mathbf{0}_{n_1 \times n_{\mathbf{L}}} \mathbf{W}]$ where $\mathbf{W} \in \mathbb{R}^{n_1 \times k}$ has i.i.d. $\mathcal{N}(0, r)$ entries. We evaluate two type of noises in this section: (1) the noise matrix \mathbf{N} has i.i.d. $\mathcal{N}(0, \sigma_{\mathbf{N}}^2)$ entries with five different values of $\sigma_{\mathbf{N}} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$; (2) \mathbf{N} has i.i.d. zero-mean Laplace entries with five difference choices of parameters $\lambda_{\mathbf{N}} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Note that \mathbf{U} , \mathbf{V} , \mathbf{W} and \mathbf{N} are mutually independent. For each $\sigma_{\mathbf{N}}$ and $\lambda_{\mathbf{N}}$, we choose $\sigma_r(\mathbf{L}) \in \{2, 4, 6, \dots\}$ and demonstrate the empirical values of $\Pr(\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}})$ (over 100 trials) in Figure 1, with and without the rescaling of $\sigma_r(\mathbf{L})$ by $\sqrt{\gamma n_2 \eta_{\mathbf{N}}}$.

In panel (a), we observe that as $\sigma_{\mathbf{N}}$ increases, the threshold of $\sigma_r(\mathbf{L})$ for correct identification of outlier columns with high probability also increases, as we expect. On the other hand, when we rescale $\sigma_r(\mathbf{L})$ by $\sqrt{\gamma n_2 \eta_{\mathbf{N}}}$ in panel (b), all curves corresponding to different values of $\sigma_{\mathbf{N}}$ are aligned together. Besides, when the ratio $\frac{\sigma_r(\mathbf{L})}{\sqrt{\gamma n_2 \eta_{\mathbf{N}}}}$ goes beyond 1, the probability of correct outlier detection is 1, which verifies our assumption (n1) in this case. Analogous results are observed for Laplace noise as well.

Next, we evaluate the bound of the column subsampling

parameter γ w.r.t. the rank r in (6). We fix \mathbf{N} as Gaussian noise with i.i.d. entries with $\sigma_{\mathbf{N}} = 0.01$ and the following discussion. Let $m = 0.3n_1$. We generate $\mathbf{L} = [\mathbf{U}\mathbf{V}^T \mathbf{0}_{n_1 \times k}]$ and $\mathbf{C} = [\mathbf{0}_{n_1 \times n_L} \mathbf{W}]$, where $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{n_L \times r}$ have i.i.d. $\mathcal{N}(0, 1)$ entries and $\mathbf{W} \in \mathbb{R}^{n_1 \times k}$ has i.i.d. $\mathcal{N}(0, r)$ entries. We choose five values of ranks $r \in \{5, 10, 15, 20, 25\}$, and plot the empirical probability of correct outlier identification $\Pr(\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}})$ versus the column subsampling parameter $\gamma \in \{0.001, 0.002, 0.003, \dots, 0.3\}$ for each r in Figure 2 (a,b). When r increases, the column subsampling parameter γ also needs to increase for correct outlier identification with high probability. If we normalize γ with $\frac{r\mu_{\mathbf{V}} \log r}{n_L}$, which is generally the dominating term in (6), then all curves corresponding to different ranks r align together, as shown in panel (b). Further, high probability of success is achieved when the ratio $\gamma / \frac{r\mu_{\mathbf{V}} \log r}{n_L} > 1$, as we have established in (6).

Analogous evaluation for the bound of the row sampling parameter m w.r.t. r in (7) is also provided. Let $\gamma = 0.2$, and the generations of \mathbf{L} , \mathbf{C} and \mathbf{N} are identical to those in the previous evaluation for γ . Again, we choose five values of ranks $r \in \{5, 10, 15, 20, 25\}$, and plot $\Pr(\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}})$ versus $m \in \{1, 2, 3, \dots, 50\}$ for each r in Figure 2. The observation matches with the bound (7) that increasing m facilitates the accurate recovery for increasing r , and the ratio $m/(r+1+\log n_2) > 1$ facilitates correct recovery with high probability, as shown in panel (c).

B. Evaluation for Incomplete Observation Settings

We proceed all experiments here using the trimming option. The setting is as follows. A trial is claimed to be a success if $\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}}$, where $\hat{\mathcal{I}}_{\mathbf{C}}$ is given in Algorithm 2 for RACOS-I. We fix $n_1 = 100$, $n_2 = 1000$, $k = 0.2n_2$ and $\lambda = 0.4$. The generations of \mathbf{L} and \mathbf{C} follow that in the previous evaluation for noisy observations. We apply the trimming step by choosing $\rho = 0.9$ throughout.

First, we evaluate the bound (16) for γ_1 w.r.t. the rank r and the entry-wise sampling parameter p respectively. In evaluating γ_1 w.r.t. r , let $\gamma_2 = 0.2$, $p = 0.5$, and the rank be chosen from $r \in \{3, 6, 9, 12, 15\}$. In evaluating γ_1 w.r.t. p , we fix $\gamma_2 = 0.2$, $r = 5$, and choose the sampling parameter from $p \in \{0.4, 0.5, 0.6, 0.7, 0.9\}$. For each p and r , we set $\gamma_1 \in \{0.02, 0.04, 0.06, \dots, 1\}$ and demonstrate the plots of $\Pr(\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}})$ versus γ_1 in Figure 3 (top row). In panel (a), we observe that when r increases, the threshold of γ_1 for correct identification with high probability also increases due to the positive dependence of γ_1 and r . Analogously in panel (c), as p increases, the threshold of γ_1 for correct identification with high probability decreases due to the inverse dependence of γ_1 on p . On the other hand, in panel (b) and (d), when we rescale γ_1 by $\frac{\mu_{\mathbf{L}} r \log(n_2)}{n_1 p}$, which is the dominating term of (16) in our setting, all curves corresponding to different values of p align gracefully and facilitates high probability of recovery with the ratio > 1 .

We carry out the similar evaluation of the bound (17) for γ_2 w.r.t. r and p respectively. We follow the same settings stated above and plot $\Pr(\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}})$ versus γ_2 in Figure 3 (bottom row). The observation is that increasing γ_2 facilitates

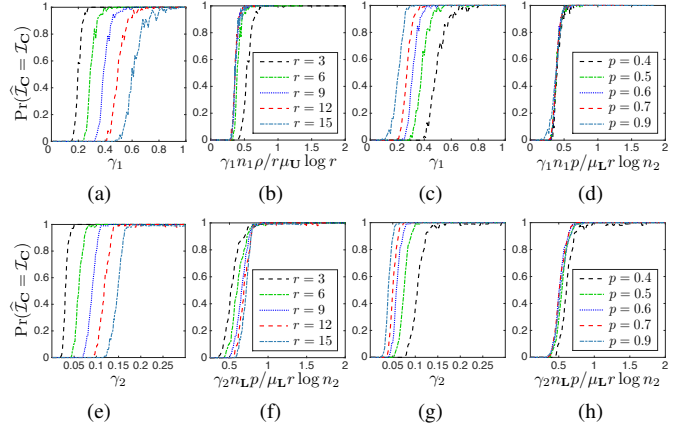


Fig. 3. Demonstration of the probability of success versus the row subsampling parameter γ_1 (top row) and the column subsampling parameter γ_2 (bottom row) under different settings of the rank r (a,b,e,f) and the entry-wise sampling parameter p (c,d,g,h). Panel (b) and (d) provide the results with rescaling of γ_1 by $\frac{\mu_{\mathbf{L}} r \log(n_2)}{n_1 p}$. Panel (f) and (h) provide the results with rescaling of γ_2 by $\frac{\mu_{\mathbf{L}} r \log(n_2)}{n_1 p}$.

the accurate recovery for increasing r and decreasing p , and the ratio $\gamma_2 / \frac{\mu_{\mathbf{L}} r \log(n_2)}{n_1 p} > 1$ corresponds to correct detection with high probability. However, we do not have explicitly $\gamma_2 = \Omega(\frac{\mu_{\mathbf{L}} r \log(n_2)}{n_1 p})$ in our bound (17), where the dominating term, considering (20), is $\gamma_2 = \Omega(\frac{\mu_{\mathbf{L}}^{3/2} r^{3/2} \log^3(n_2)}{n_1 p})$. This suggests that further improvement may be achieved in terms of the sampling complexity of γ_2 in (20), which we leave for future investigation.

C. Timing Performance

We further examine the timing performances for both models. We fix $n_1 = 500$, $n_2 = 1000$, $k = 0.2n_2$, $n_L = n_2 - k$, and $\lambda = 0.4$, and generate \mathbf{L} , \mathbf{C} , and the Gaussian noise \mathbf{N} in the same way described above.

For the noisy observation setting, we fix $r = 10$ and choose different combinations of the row sampling parameter m and the column sampling parameter γ . More specifically, we choose $m \in \{10, 20, 30, \dots, 500\}$ and $\gamma \in \{0.02, 0.04, 0.06, \dots, 1\}$, where the pair $(m, \gamma) = (500, 1)$ corresponds to operating on the full-size data matrix \mathbf{M} . We first provide the “phase transition” behavior as discussed in [15] for all combinations of m and γ and a fixed $\lambda = 0.5$ in OP. Then we record the CPU execution time of Algorithm 1. The phase transition and the contour plot of timing evaluation are provided in Figure 4 (a,b). The values on contour lines are the speed-ups of algorithm compared with the full size model, *i.e.* $(m, \gamma) = (500, 1)$. We can see that our approach shows significant advantage in terms of computational efficiency over the full data model when m and γ are small. For example, when $(m/n_1, \gamma) = (0.1, 0.1)$, our approach is > 100 times faster than that using the full data. Another interesting observation is that the full size model $(m, \gamma) = (500, 1)$ is not the slowest here, while the nearly full size model is the slowest. This is because in the full data model, we do not need to construct the random projection matrices and the corresponding projection operations. In applications, such as

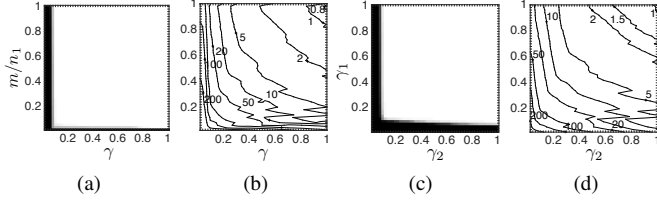


Fig. 4. Demonstration of the performance using different combinations of m and γ for noisy observations and different combinations of γ_1 and γ_2 for incomplete observations via (a,c) phase transition and (b,d) timing evaluation of OP/MP respectively.

the salient image feature detection, speedup of over 100 times can be achieved with comparable performances [15], [35].

Analogous evaluation is also carried out for the incomplete observation setting. We fix $p = 0.4$ and $r = 5$, and choose different combinations of the row sampling parameter $\gamma_1 \in \{0.02, 0.04, 0.06, \dots, 1\}$ and the column sampling parameter $\gamma_2 \in \{0.02, 0.04, 0.06, \dots, 1\}$, where the pair $(\gamma_1, \gamma_2) = (1, 1)$ corresponds to the full-size data model. We provide the phase transition and the contour plot of timing evaluation of Algorithm 2 for each pair of (γ_1, γ_2) in Figure 4 (c,d). Significant improvement of the computational efficiency over the full data model is also observed. For example, when $(\gamma_1, \gamma_2) = (0.2, 0.2)$, our approach is > 50 times faster than the full data model.

V. DISCUSSION

The idea of identifying outliers from a few linear summaries of the original data matrix may be extended to a large class of models. The key insight is that if the structure of the problem can be (approximately) preserved, then significantly improved computational complexity may be achieved by operating on a much smaller dimensional problem. This is also closely related with the recent development of sketching techniques in linear algebra, data mining, and machine learning [37]. Our future interest includes, but not limited to, extension of a more challenging observation model in the existence of both missing entries and noise, and the identification of outliers from union of subspaces.

VI. APPENDIX

A. Proof of Lemma III.1

We start with demonstrating (23), ($\tilde{n}1$) and ($\tilde{n}2$). We state clearly here that all following results are obtained by taking $\varepsilon = \sqrt{2}/4$. The choice of $\sqrt{2}/4$ is somewhat arbitrary and we choose this fixed value for concreteness. Note that given $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$, if Φ satisfies the distributional JL property with m specified in (7), then Φ is an ε -stable embedding of $(\mathcal{L}, \cup_{i \in \mathcal{I}_C} \{\mathbf{C}_{:,i} + \mathbf{N}_{:,i}\} \cup_{j \in [n_2]} \{\mathbf{N}_{:,j}\} \cup \{\mathbf{0}\})$ with probability at least $1 - \delta$ (Lemma III.1 in [15]). This implies $\sqrt{1 - \varepsilon} \|\mathbf{N}_{:,i}\|_2 \leq \|\tilde{\mathbf{N}}_{:,i}\|_2 \leq \sqrt{1 + \varepsilon} \|\mathbf{N}_{:,i}\|_2$ for any $i \in [n_2]$, which results in (23).

We also have,

$$\sigma_r(\tilde{\mathbf{L}}) \stackrel{(i)}{\geq} \sqrt{1 - \varepsilon} \sigma_r(\mathbf{L}) \stackrel{(ii)}{>} \sqrt{1 - \varepsilon} \frac{90\sqrt{2}\gamma}{\tau_1} n_2 \eta_{\mathbf{N}} \stackrel{(iii)}{\geq} \frac{72\sqrt{2}\gamma}{\tau_1} n_2 \eta_{\mathbf{N}},$$

where (i) is a direct application of Theorem 1 in [38] via the ε -stable embedding property, (ii) is from ($\mathbf{n}1$), and (iii) is from (23), which results in ($\tilde{n}1$).

To verify ($\tilde{n}2$), we have from the ε -stable embedding property of Φ and ($\mathbf{n}2$) that for any $i \in \mathcal{I}_C$, $j \in [n_2]$,

$$\|\tilde{\mathbf{C}}_{:,i}\|_2 \geq \sqrt{1 - \varepsilon} \|\mathbf{C}_{:,i}\|_2 > \sqrt{1 - \varepsilon} \tau_2 \|\mathbf{N}_{:,j}\|_2.$$

Next, we demonstrate ($\tilde{d}1$)-($\tilde{d}5$). ($\tilde{d}1$)-($\tilde{d}3$) follow directly from the result in [15] (Lemma III.1). We have from Theorem 1 in [38],

$$\frac{\sigma_1(\tilde{\mathbf{L}})}{\sigma_r(\tilde{\mathbf{L}})} \leq \frac{\sqrt{1 + \varepsilon} \sigma_1(\mathbf{L})}{\sqrt{1 - \varepsilon} \sigma_r(\mathbf{L})} \leq \sqrt{3} \kappa, \quad (25)$$

which establish ($\tilde{d}4$). For ($\tilde{d}5$), we have

$$\begin{aligned} \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{C}}_{:,i}\|_2 &\stackrel{(i)}{\geq} \sqrt{\frac{1 - 2\varepsilon}{1 - \varepsilon}} \|\mathbf{P}_{\mathcal{L}^\perp} \mathbf{C}_{:,i}\|_2 \stackrel{(ii)}{>} \tau_1 \sqrt{\frac{1 - 2\varepsilon}{1 - \varepsilon}} \|\mathbf{C}_{:,i}\|_2 \\ &\stackrel{(iii)}{\geq} \tau_1 \sqrt{\frac{1 - 2\varepsilon}{1 - \varepsilon^2}} \|\tilde{\mathbf{C}}_{:,i}\|_2 \geq \tau_1 \|\tilde{\mathbf{C}}_{:,i}\|_2 / 2, \end{aligned}$$

where (i) and (iii) are from Φ being an ε -stable embedding, and (ii) is from ($\mathbf{d}5$). It is straightforward from the definition of the orthogonal projection $\mathbf{P}_{\tilde{\mathcal{L}}}$ that for any $i \in \mathcal{I}_C$ and $j \in \mathcal{I}_L$,

$$\begin{aligned} \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} (\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2 &= \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{N}}_{:,j}\|_2 < \|\tilde{\mathbf{N}}_{:,j}\|_2 \\ &\stackrel{(i)}{<} \sqrt{3} \|\tilde{\mathbf{C}}_{:,i}\|_2 / \tau_2 \stackrel{(ii)}{<} \tau_1 \|\tilde{\mathbf{C}}_{:,i}\|_2 / 2, \end{aligned}$$

where (i) is from (23) and ($\tilde{n}2$), and (ii) is from the fact $\tau_2 > 2\sqrt{3}/\tau_1$ from (9). By definition of $\mathcal{I}_{\tilde{\mathcal{C}}}$, we have $\mathcal{I}_{\tilde{\mathcal{C}}} = \mathcal{I}_C$.

B. Proof of Lemma III.2

For notional convenience, we introduce

$$\tilde{\mathbf{M}} \triangleq \tilde{\mathbf{M}}\mathbf{S} = \tilde{\mathbf{L}}\mathbf{S} + \tilde{\mathbf{C}}\mathbf{S} + \tilde{\mathbf{N}}\mathbf{S} = \tilde{\mathbf{L}} + \tilde{\mathbf{C}} + \tilde{\mathbf{N}},$$

where \mathbf{S} is the column sampling matrix. From Lemma III.2 in [15], the following results hold with probability at least $1 - \delta$ when \mathbf{S} is generated as specified²:

- (a1) \mathbf{S} has $(1/2)\gamma n_2 \leq \tilde{n}_2 \leq (3/2)\gamma n_2$ columns,
- (a2) $\tilde{\mathbf{L}}$ has $\tilde{n}_L \leq (3/2)\gamma n_L$ nonzero columns,
- (a3) $\tilde{\mathbf{C}}$ has $\tilde{k} \leq (3/2)\gamma k_u$ nonzero columns,
- (a4) $\sigma_1^2(\tilde{\mathbf{V}}^* \mathbf{S}) \leq (3/2)\gamma$, and
- (a5) $\sigma_r^2(\tilde{\mathbf{V}}^* \mathbf{S}) \geq (1/2)\gamma$,

where $\tilde{\mathbf{V}}$ is the matrix of right singular vectors from the compact SVD of $\tilde{\mathbf{L}}$, i.e. $\tilde{\mathbf{L}} = \tilde{\mathbf{U}} \tilde{\Sigma} \tilde{\mathbf{V}}^*$, and $\sigma_i(\tilde{\mathbf{V}}^* \mathbf{S})$ denotes the i -th largest singular value of $\tilde{\mathbf{V}}^* \mathbf{S}$. Note that parameters (1/2) and (3/2) arising in the conditions (a1)-(a5) are somewhat arbitrary, and they are fixed to the values here for ease of exposition.

Claim (I) follows directly from (a1). To justify Claim (ii), we have (with a minor modification of Lemma III.2 in [15]) that when (a1)-(a5) and (6) are satisfied, the following structural conditions of $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{C}}$ hold:

- ($\tilde{d}1$) $\tilde{r} = \text{rank}(\tilde{\mathbf{L}}) = r$,
- ($\tilde{d}2$) $\tilde{\mathbf{L}}$ has $n_{\tilde{\mathbf{L}}} \leq \frac{3}{2}\gamma n_L$ nonzero columns,
- ($\tilde{d}3$) $\tilde{\mathbf{L}}$ satisfies the *column incoherence property* with parameter $\mu_{\tilde{\mathbf{V}}} = 9\mu_{\mathbf{V}}$, and
- ($\tilde{d}4$) $\mathcal{I}_{\tilde{\mathcal{C}}} \triangleq \{i : \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} (\tilde{\mathbf{C}}_{:,i})\|_2 > 0\}$ with $|\mathcal{I}_{\tilde{\mathcal{C}}}| = \tilde{k}$, where $\tilde{\mathcal{L}}$ denotes the linear subspace spanned by columns of $\tilde{\mathbf{L}}$

²Here we use $\delta/6$ instead of $\delta/5$ for bounding the probability of the complement of each event (a1)-(a5) in the proof of Lemma III.2 of [15], and a different (a3) with that in [15].

and $\mathbf{P}_{\tilde{\mathcal{L}}^\perp}$ is the orthogonal projection operator onto the orthogonal complement of $\tilde{\mathcal{L}}$ in \mathbb{R}^m , and

$$\tilde{k} \leq \left(\frac{1}{1 + (1024/9) \tilde{r} \mu_{\tilde{\mathbf{V}}}} \right) \tilde{n}_2. \quad (26)$$

Proofs of (d1)-(d3) are identical to those in Lemma A.4 in [15]. The condition (d4) follows from (a3) since

$$\tilde{k} \leq \frac{3}{2} \gamma k_u \stackrel{(i)}{\leq} \frac{3k_u \tilde{n}_2}{n_2} \stackrel{(ii)}{=} \left(\frac{1}{1 + (1024/9) \tilde{r} \mu_{\tilde{\mathbf{V}}}} \right) \tilde{n}_2,$$

where (i) is from (a1), and (ii) is from (d1) and (d3).

Now we verify (a3), which will be discussed in two cases.

Case 1. Since $\gamma \geq \frac{600(1+1024r\mu_{\tilde{\mathbf{L}}}) \log(\frac{6}{\delta})}{n_2}$ in (6), we have $\frac{200}{\gamma} \log(\frac{6}{\delta}) \leq \frac{n_2}{3(1+1024r\mu_{\tilde{\mathbf{V}}})}$. Let k satisfy

$$\frac{200}{\gamma} \log(\frac{6}{\delta}) \leq k \leq k_u = \frac{n_2}{3(1 + 1024r\mu_{\tilde{\mathbf{V}}})}. \quad (27)$$

Note that \tilde{k} is a Hypergeometric random variable with distributions $\text{Hyp}(n_2, \tilde{n}_2, k)$, which is parameterized by the population size n_2 , the total number of draws \tilde{n}_2 , and the total positive elements k . Then we have from [15] (second part of the proof of Lemma III.2) that

$$\Pr(\tilde{k} > (3/2)\gamma k) \leq \exp(-\gamma k/200). \quad (28)$$

Let the R.H.S. of (28) be no larger than $\delta/6$. Then we have $\tilde{k} \leq (3/2)\gamma k \leq (3/2)\gamma k_u$, i.e., (a3) holds, w.p. $\geq 1 - 6/\delta$, if k satisfies (27).

Case 2. If $k < \frac{200}{\gamma} \log(\frac{6}{\delta})$, then (a3) holds w.p. $\geq 1 - 6/\delta$ by the stochastic ordering argument [39]. More specifically, let \tilde{k}_1 and \tilde{k}_2 be Hypergeometric random variables with distributions $\text{Hyp}(n_2, \tilde{n}_2, k_1)$ and $\text{Hyp}(n_2, \tilde{n}_2, k_2)$ respectively, where $k_1 > k_2$. Then by Lemma 4.1 in [16], for any $x \in [0, \infty)$,

$$\Pr(\tilde{k}_2 \leq x) \geq \Pr(\tilde{k}_1 \leq x).$$

This indicates that when $k < \frac{200}{\gamma} \log(\frac{6}{\delta})$, we have $\tilde{k} \leq (3/2)\gamma k_u$ w.p. $\geq 1 - 6/\delta$.

Next, we show that the estimate $\hat{\mathbf{L}}$ of $\tilde{\mathbf{L}}$ can be obtained with the existence of noise, and $\|\tilde{\mathbf{L}} - \hat{\mathbf{L}}\|_2$ can be bounded in terms of $\|\tilde{\mathbf{N}}\|_{1,2}$. We formalize this notion in Lemma VI.1, and provide the proof in Appendix VI-G.

Lemma VI.1 (Outlier Pursuit with Noise, adapted from Theorem 2 of [9]). *Let $\tilde{\mathbf{M}} = \tilde{\mathbf{L}} + \tilde{\mathbf{C}} + \tilde{\mathbf{N}}$ be an $m \times \tilde{n}_2$ matrix whose components $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{C}}$ satisfy the structural conditions (d1)-(d4) with \tilde{k} satisfying (26). Then for $\lambda = \frac{\sqrt{9+1024\mu_{\tilde{\mathbf{L}}}r}}{14\sqrt{\tilde{n}_2}}$ and any solution pair obtained from the outlier pursuit*

$$\begin{aligned} \{\hat{\mathbf{L}}, \hat{\mathbf{C}}\} &= \underset{\mathbf{L}, \mathbf{C}}{\operatorname{argmin}} \|\mathbf{L}\|_* + \lambda \|\mathbf{C}\|_{1,2} \\ \text{s.t. } &\|\tilde{\mathbf{M}} - (\mathbf{L} + \mathbf{C})\|_F \leq \varepsilon_1, \end{aligned} \quad (29)$$

there exists $\tilde{\mathbf{L}}_0$ and $\tilde{\mathbf{C}}_0$ such that $\tilde{\mathbf{M}}_0 = \tilde{\mathbf{L}}_0 + \tilde{\mathbf{C}}_0$, where $\tilde{\mathbf{L}}_0$ has the correct column space of $\tilde{\mathbf{L}}$, $\tilde{\mathbf{C}}_0$ has the correct column support of $\tilde{\mathbf{C}}$, and

$$\|\tilde{\mathbf{L}}_0 - \hat{\mathbf{L}}\|_2 \leq 10\|\tilde{\mathbf{N}}\|_{1,2}, \quad \|\tilde{\mathbf{C}}_0 - \hat{\mathbf{C}}\|_2 \leq 9\|\tilde{\mathbf{N}}\|_{1,2}. \quad (30)$$

Note that we can only guarantee the estimation errors in terms of some $\tilde{\mathbf{L}}_0$ having the same column space with $\tilde{\mathbf{L}}$ and some $\tilde{\mathbf{C}}_0$ sharing the same column support with $\tilde{\mathbf{C}}$, which is

enough for our purpose of analysis. Lemma VI.1 is further utilized to bound $\|\mathbf{P}_{\hat{\mathcal{L}}(1)} - \mathbf{P}_{\tilde{\mathcal{L}}}\|_2$ away from 1, where $\hat{\mathcal{L}}(1)$ is the column space of $\hat{\mathbf{L}}(1)$, obtained by the singular values thresholding operation of $\hat{\mathbf{L}}$. We will show that $\hat{\mathbf{L}}(1)$ has the same rank with $\tilde{\mathbf{L}}$ from our choices of parameters.

Remind that $\tilde{\mathbf{L}}$ has zero columns when the corresponding columns of $\tilde{\mathbf{C}}$ are nonzero. Since $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{C}}_0$ have the same column support, thus $\tilde{\mathbf{L}}_0$ and $\tilde{\mathbf{L}}$ are identical for non-zero columns of $\tilde{\mathbf{L}}$. Besides, columns of $\tilde{\mathbf{L}}_0$ may be non-zero at the locations of zero columns of $\tilde{\mathbf{L}}$. Therefore, we have that for any $i \in [r]$, $\sigma_i(\tilde{\mathbf{L}}_0) \geq \sigma_i(\tilde{\mathbf{L}})$. This can be seen using the following argument. Since $\tilde{\mathbf{L}}_0$ and $\tilde{\mathbf{L}}$ have the same column space, then the i -th singular values of $\tilde{\mathbf{L}}_0$ and $\tilde{\mathbf{L}}$ satisfy

$$\sigma_i(\tilde{\mathbf{L}}_0) = \|\mathbf{u}_i^T \tilde{\mathbf{L}}_0\|_2 \geq \|\mathbf{u}_i^T \tilde{\mathbf{L}}\|_2 = \sigma_i(\tilde{\mathbf{L}}), \quad (31)$$

where \mathbf{u}_i is the i -th left singular value of $\tilde{\mathbf{L}}_0$. Combining (n1), (a5), and (31), we have

$$\sigma_r(\tilde{\mathbf{L}}_0) \geq \sigma_r(\tilde{\mathbf{L}}) \geq \sqrt{\frac{\gamma}{2}} \sigma_r(\tilde{\mathbf{L}}) \geq \frac{72}{\tau_1} \gamma n_2 \eta_{\mathbf{N}}. \quad (32)$$

On the other hand, we have

$$\begin{aligned} \sigma_r(\tilde{\mathbf{L}}_0) - \sigma_r(\hat{\mathbf{L}}) &\stackrel{(i)}{\leq} \|\tilde{\mathbf{L}}_0 - \hat{\mathbf{L}}\|_2 \stackrel{(ii)}{\leq} 10\|\tilde{\mathbf{N}}\|_{1,2} \stackrel{(iii)}{\leq} 12\tilde{n}_2 \eta_{\mathbf{N}} \\ &\stackrel{(iv)}{\leq} \frac{12 \cdot 3\gamma n_2 \eta_{\mathbf{N}}}{2} = 18\gamma n_2 \eta_{\mathbf{N}}, \end{aligned} \quad (33)$$

where (i) is from Lemma III.1 in the supplemental material, (ii) is from (30), (iii) is from (23), and (iv) is from (a1). Combining (32), (33), and $\tau_1 \in (0, 1)$, we have

$$\sigma_r(\hat{\mathbf{L}}) \geq \left(\frac{4}{\tau_1} - 1 \right) 18\gamma n_2 \eta_{\mathbf{N}} \geq 54\gamma n_2 \eta_{\mathbf{N}} > 0. \quad (34)$$

Using the same argument as (33) and (34), we have

$$\sigma_{r+1}(\hat{\mathbf{L}}) \leq \sigma_{r+1}(\tilde{\mathbf{L}}_0) + \|\tilde{\mathbf{L}}_0 - \hat{\mathbf{L}}\|_2 \leq 18\gamma n_2 \eta_{\mathbf{N}}. \quad (35)$$

When the singular value thresholding constant α satisfies (11), it is guaranteed by (34) and (35) that $\text{rank}(\hat{\mathbf{L}}(1)) = r$, where $\hat{\mathbf{L}}(1) = \hat{\mathbf{U}} \mathcal{D}_\alpha(\hat{\mathbf{\Sigma}}) \hat{\mathbf{V}}^*$ and $\mathcal{D}_\alpha(\hat{\mathbf{\Sigma}})$ is the hard thresholding operation. Combining (32) and (33), we have

$$\begin{aligned} \|\mathbf{P}_{\hat{\mathcal{L}}(1)} - \mathbf{P}_{\tilde{\mathcal{L}}}\|_2 &\stackrel{(i)}{=} \|\mathbf{P}_{\hat{\mathcal{L}}(1)} - \mathbf{P}_{\tilde{\mathcal{L}}}\|_2 \stackrel{(ii)}{\leq} \frac{\|\tilde{\mathbf{L}} - \hat{\mathbf{L}}(1)\|_2}{\sigma_r(\tilde{\mathbf{L}})} \\ &\stackrel{(iii)}{\leq} \frac{\|\tilde{\mathbf{L}} - \hat{\mathbf{L}}\|_2}{\sigma_r(\tilde{\mathbf{L}})} \stackrel{(iv)}{\leq} \frac{18\gamma n_2 \eta_{\mathbf{N}}}{\sigma_r(\tilde{\mathbf{L}})}, \end{aligned} \quad (36)$$

where (i) is from $\mathbf{P}_{\tilde{\mathcal{L}}} = \mathbf{P}_{\hat{\mathcal{L}}}$, (ii) is from the additive perturbation bound of the orthogonal projection in Lemma III.2 in the supplemental material, (iii) is from the way we generate $\hat{\mathbf{L}}(1)$, and (iv) is from (33). Let τ_3 be the R.H.S. of (36). Combining (32) and (36), we have

$$\|\mathbf{P}_{\hat{\mathcal{L}}(1)} - \mathbf{P}_{\tilde{\mathcal{L}}}\|_2 \leq \tau_3 \leq \frac{18\gamma n_2 \eta_{\mathbf{N}}}{72\gamma n_2 \eta_{\mathbf{N}}/\tau_1} = \frac{\tau_1}{4}, \quad (37)$$

Using triangle inequality and Cauchy-Schwarz inequality, we have for any $j \in \mathcal{I}_{\tilde{\mathbf{L}}}$

$$\begin{aligned} &\left| \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp}(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2 - \|\mathbf{P}_{\hat{\mathcal{L}}^\perp}(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2 \right| \\ &\leq \|(\mathbf{P}_{\tilde{\mathcal{L}}^\perp} - \mathbf{P}_{\hat{\mathcal{L}}^\perp})(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2 \\ &\leq \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} - \mathbf{P}_{\hat{\mathcal{L}}^\perp}\|_2 \cdot \|(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2 \\ &= \|\mathbf{P}_{\hat{\mathcal{L}}(1)} - \mathbf{P}_{\tilde{\mathcal{L}}}\|_2 \cdot \|(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2 \leq \tau_3 \|(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2, \end{aligned}$$

from which we have for any $j \in \mathcal{I}_L$ and $i \in \mathcal{I}_C$,

$$\begin{aligned} \|\mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2 &\leq \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp}(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2 + \tau_3 \|\tilde{\mathbf{L}} + \tilde{\mathbf{N}}\|_{:,j} \\ &= \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{N}}_{:,j}\|_2 + \tau_3 \|\tilde{\mathbf{L}} + \tilde{\mathbf{N}}\|_{:,j}. \end{aligned} \quad (38)$$

Applying the same analysis, we have for any $i \in \mathcal{I}_C$

$$\begin{aligned} \|\mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{C}} + \tilde{\mathbf{N}})_{:,i}\|_2 \\ \geq \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp}(\tilde{\mathbf{C}} + \tilde{\mathbf{N}})_{:,i}\|_2 - \tau_3 \|\tilde{\mathbf{C}} + \tilde{\mathbf{N}}\|_{:,i}. \end{aligned} \quad (39)$$

Then for any $j \in \mathcal{I}_L$ and $i \in \mathcal{I}_C$, we have

$$\begin{aligned} \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp}(\tilde{\mathbf{C}}_{:,i})\|_2 &\stackrel{(i)}{>} \frac{\tau_1}{2} \|\tilde{\mathbf{C}}_{:,i}\|_2 \stackrel{(ii)}{>} \frac{\tau_1}{4} \left(\|\tilde{\mathbf{C}}_{:,i}\|_2 + \frac{4}{5} \tau_2 \eta_N \right) \\ &\stackrel{(iii)}{>} \frac{\tau_1}{4} \|\tilde{\mathbf{C}}_{:,i}\|_2 + \left(\frac{6}{5} (\beta + 1) \left(\frac{\tau_1}{4} + 1 \right) + 18\sqrt{6}\gamma\beta\kappa n_2 \right) \eta_N \\ &\stackrel{(iv)}{\geq} \frac{\tau_1}{4} \|\tilde{\mathbf{C}}_{:,i}\|_2 + (\beta + 1) \left(\frac{\tau_1}{4} + 1 \right) \eta_N + 18\sqrt{6}\gamma\beta\kappa n_2 \eta_N \\ &\stackrel{(v)}{\geq} \frac{\tau_1}{4} \|\tilde{\mathbf{C}}_{:,i}\|_2 + \left(\frac{\tau_1}{4} + 1 \right) \|\tilde{\mathbf{N}}_{:,i}\|_2 + \beta \left(\frac{\tau_1}{4} + 1 \right) \|\tilde{\mathbf{N}}_{:,j}\|_2 \\ &\quad + 18\sqrt{6}\gamma\beta\kappa n_2 \eta_N \\ &\stackrel{(vi)}{\geq} \frac{\tau_1}{4} \|\tilde{\mathbf{C}}_{:,i}\|_2 + \frac{\tau_1}{4} \|\tilde{\mathbf{N}}_{:,i}\|_2 + \beta \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{N}}_{:,j}\|_2 + \frac{\beta\tau_1}{4} \|\tilde{\mathbf{N}}_{:,j}\|_2 \\ &\quad + 18\sqrt{6}\gamma\beta\kappa n_2 \eta_N + \|\tilde{\mathbf{N}}_{:,i}\|_2, \end{aligned} \quad (40)$$

where (i) is from (d5), (ii) is from (n2), (iii) is from (9), (iv) is from (23), (v) is from the definition of η_N , and (vi) is from the condition $\beta > 1$. Then, we have

$$\begin{aligned} \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp}(\tilde{\mathbf{C}}_{:,i} + \tilde{\mathbf{N}}_{:,i})\|_2 &\geq \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp}(\tilde{\mathbf{C}}_{:,i})\|_2 - \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp}(\tilde{\mathbf{N}}_{:,i})\|_2 \\ &\stackrel{(i)}{>} \frac{\tau_1}{4} \|\tilde{\mathbf{C}}_{:,i}\|_2 + \frac{\tau_1}{4} \|\tilde{\mathbf{N}}_{:,i}\|_2 + \beta \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{N}}_{:,j}\|_2 + \frac{\beta\tau_1}{4} \|\tilde{\mathbf{N}}_{:,j}\|_2 \\ &\quad + 18\sqrt{6}\gamma\beta\kappa n_2 \eta_N \\ &\stackrel{(ii)}{\geq} \frac{\tau_1}{4} \|\tilde{\mathbf{C}} + \tilde{\mathbf{N}}\|_{:,i} + \beta \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{N}}_{:,j}\|_2 \\ &\quad + 18\sqrt{\gamma}\beta n_2 \eta_N \left(\sqrt{6}\kappa + \frac{\sqrt{\gamma} \|\tilde{\mathbf{N}}_{:,j}\|_2}{\sigma_r(\tilde{\mathbf{L}})} \right) \\ &\stackrel{(iii)}{\geq} \frac{\tau_1}{4} \|\tilde{\mathbf{C}} + \tilde{\mathbf{N}}\|_{:,i} + \beta \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{N}}_{:,j}\|_2 \\ &\quad + 18\sqrt{\gamma}\beta n_2 \eta_N \left(\frac{\sqrt{2}\sigma_1(\tilde{\mathbf{L}})}{\sigma_r(\tilde{\mathbf{L}})} + \frac{\sqrt{\gamma} \|\tilde{\mathbf{N}}_{:,j}\|_2}{\sigma_r(\tilde{\mathbf{L}})} \right) \\ &\stackrel{(iv)}{\geq} \frac{\tau_1}{4} \|\tilde{\mathbf{C}} + \tilde{\mathbf{N}}\|_{:,i} + \beta \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{N}}_{:,j}\|_2 \\ &\quad + \beta \frac{18\gamma n_2 \eta_N}{\sigma_r(\tilde{\mathbf{L}})} (\|\tilde{\mathbf{L}}_{:,j}\|_2 + \|\tilde{\mathbf{N}}_{:,j}\|_2) \\ &\stackrel{(v)}{=} \frac{\tau_1}{4} \|\tilde{\mathbf{C}} + \tilde{\mathbf{N}}\|_{:,i} + \beta \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{N}}_{:,j}\|_2 \\ &\quad + \beta\tau_3 (\|\tilde{\mathbf{L}}_{:,j}\|_2 + \|\tilde{\mathbf{N}}_{:,j}\|_2) \\ &\stackrel{(vi)}{\geq} \tau_3 \|\tilde{\mathbf{C}} + \tilde{\mathbf{N}}\|_{:,i} + \beta(\tau_3 \|\tilde{\mathbf{L}} + \tilde{\mathbf{N}}\|_{:,j} + \|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{N}}_{:,j}\|_2), \end{aligned} \quad (41)$$

where (i) is from (40), (ii) is from (32), (iii) is from (d4), (iv) is from (32) and fact that $\max_{j \in [n_2]} \|\tilde{\mathbf{L}}_{:,j}\|_2 \leq \sigma_1(\tilde{\mathbf{L}})$, (v) is from (36), and (vi) is from (37). Combining (38), (39) and (41), we have

$$\begin{aligned} \|\mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{C}} + \tilde{\mathbf{N}})_{:,j}\|_2 &> \beta \left(\|\mathbf{P}_{\tilde{\mathcal{L}}^\perp} \tilde{\mathbf{N}}_{:,j}\|_2 + \tau_3 \|\tilde{\mathbf{L}} + \tilde{\mathbf{N}}\|_{:,j} \right) \\ &\geq \beta \|\mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2. \end{aligned}$$

Then Claim (II) of Theorem II.1 is verified.

C. Proof of Lemma III.3

We first show that when q satisfies (8), the random projection via $\Psi \in \mathbb{R}^{q \times m}$ approximately preserves (24) with high probability. This is formalized in the following lemma.

Lemma VI.2. Suppose $\Psi \in \mathbb{R}^{q \times m}$ is drawn from a distribution satisfying the distributional JL property (4) with q satisfying (8), and (24) holds. Given $\delta \in (0, 1)$, then for any $i \in \mathcal{I}_C$ and $j \in \mathcal{I}_L$, with probability at least $1 - \delta$, we have

$$\|\Psi \mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{C}} + \tilde{\mathbf{N}})_{:,i}\|_2 > \frac{\sqrt{3}}{3} \beta \|\Psi \mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2. \quad (42)$$

Proof. For any $\varepsilon \in (0, 1)$, if Ψ is as specified and (24) holds, then with probability at least $1 - \delta$, we have

$$\begin{aligned} \|\Psi \mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{C}} + \tilde{\mathbf{N}})_{:,i}\|_2 &\geq \sqrt{\frac{1-\varepsilon}{1+\varepsilon}} \|\mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{C}} + \tilde{\mathbf{N}})_{:,i}\|_2 \\ &> \sqrt{\frac{1-\varepsilon}{1+\varepsilon}} \beta \|\mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2 \\ &\geq \sqrt{\frac{1-\varepsilon}{1+\varepsilon}} \beta \|\Psi \mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2. \end{aligned}$$

Then (42) holds if we take $\varepsilon = \sqrt{2}/4$. \square

Now it is straightforward to see that if $\beta > \sqrt{3}$ and we choose some ε_2 that satisfies $\max_{j \in \mathcal{I}_L} \|\Psi \mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{L}} + \tilde{\mathbf{N}})_{:,j}\|_2 < \varepsilon_2 < \min_{i \in \mathcal{I}_C} \|\Psi \mathbf{P}_{\tilde{\mathcal{L}}_{(1)}^\perp}(\tilde{\mathbf{C}} + \tilde{\mathbf{N}})_{:,i}\|_2$, then we have $\hat{\mathcal{I}}_C = \mathcal{I}_{\tilde{\mathbf{C}}}$.

D. Proof of Theorem II.2

We only need to bound the constants η_N , which further implies the bound of α . The rest of the proof is identical to that of Theorem II.1. Since \mathbf{N} has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, then for any $i \in [n_2]$, $\|\mathbf{N}_{:,i}\|_2^2 / \sigma^2 = \sum_{j=1}^{n_1} (N_{ji} / \sigma)^2$ has chi-square distribution $\chi_{n_1}^2$ with n_1 degree of freedom. Given $t \in (0, 1)$, we have the following tail bounds [40],

$$\begin{aligned} \mathcal{P}(\|\mathbf{N}_{:,i}\|_2^2 \geq \sigma^2 n_1 (1+t)) &\leq e^{-n_1 t^2/8} \quad \text{and} \\ \mathcal{P}(\|\mathbf{N}_{:,i}\|_2^2 \leq \sigma^2 n_1 (1-t)) &\leq e^{-n_1 t^2/8}. \end{aligned}$$

Let $t = \sqrt{s/n_1}$ for some $s \in (0, n_1)$, then we have

$$\begin{aligned} \mathcal{P}\left(\|\mathbf{N}_{:,i}\|_2 \geq \sigma \sqrt{n_1 + \sqrt{n_1 s}}\right) &\leq e^{-s/8} \quad \text{and} \\ \mathcal{P}\left(\|\mathbf{N}_{:,i}\|_2 \leq \sigma \sqrt{n_1 - \sqrt{n_1 s}}\right) &\leq e^{-s/8}. \end{aligned}$$

By union bound, we further have

$$\begin{aligned} \mathcal{P}\left(\max_{i \in [n_2]} \|\mathbf{N}_{:,i}\|_2 \leq \sigma \sqrt{n_1 - \sqrt{n_1 s}}\right) &\leq n_2 e^{-s/8} \quad \text{and} \\ \mathcal{P}\left(\max_{i \in [n_2]} \|\mathbf{N}_{:,i}\|_2 \geq \sigma \sqrt{n_1 + \sqrt{n_1 s}}\right) &\leq n_2 e^{-s/8}, \end{aligned}$$

Let $\delta = 2n_2 e^{-s/8} \in (0, 1)$ and apply the union bound, then with probability at least $1 - \delta$, we have

$$C_1 \sigma \leq \eta_N \leq C_2 \sigma, \quad (43)$$

where C_1 and C_2 are specified as in Theorem II.2.

Combining (13) and (43), we have α satisfies (11). Finally, the overall results of Theorem II.2 hold via the union bound.

E. Proof of Theorem II.3

We follow the idea of the proof for Theorem II.1 by providing the proof sketch for Theorem II.3, which is formalized by the intermediate results Lemma VI.3 and Lemma VI.4. The proofs of the lemmata are provided later.

For notional convenience, we introduce:

$$\tilde{\mathbf{M}} = \Phi \mathbf{M}, \tilde{\mathbf{L}} = \Phi \mathbf{L}, \tilde{\mathbf{C}} = \Phi \mathbf{C}, \text{ and } \tilde{\mathbf{C}}_\Omega = \Phi \mathbf{P}_\Omega(\mathbf{C}).$$

We start with showing that the analogous structural conditions for $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{C}}$ also hold for $\tilde{\mathbf{M}}$ provided that the row sampling parameter γ_1 is sufficiently large. This is stated in Lemma VI.3, and we provide the proof in Section VI-E1.

Lemma VI.3. Suppose matrices $\mathbf{L}, \mathbf{C} \in \mathbb{R}^{n_1 \times n_2}$ satisfy the structural conditions (g1)-(g4) with p satisfying (14). Given $\delta \in (0, 1)$, further suppose $\Phi \in \mathbb{R}^{m \times n_1}$ is a row sampling matrix with the sampling parameter γ_1 satisfying (16). Then, with probability at least $1 - \delta$, the components $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{C}}$ satisfy

- (g1) $\text{rank}(\tilde{\mathbf{L}}) = r$,
- (g2) $\tilde{\mathbf{L}}$ has $n_{\mathbf{L}}$ nonzero columns,
- (g3) $\tilde{\mathbf{L}}$ satisfies the row and column incoherence property with parameters $\mu_{\tilde{\mathbf{U}}} = 9\mu_{\mathbf{U}}$ and $\mu_{\tilde{\mathbf{V}}} = \mu_{\mathbf{V}}$ respectively, and
- (g4) $\mathcal{I}_{\tilde{\mathbf{C}}} \triangleq \{j \in [n_2] : \|(\mathbf{P}_{\tilde{\mathbf{L}}^\perp}(\tilde{\mathbf{C}}_\Omega))_{:,j}\|_2 > 0\} = \mathcal{I}_{\mathbf{C}}$, where $\tilde{\mathcal{L}}$ denotes the subspace spanned by columns of $\tilde{\mathbf{L}}$, and $\mathbf{P}_{\tilde{\mathbf{L}}^\perp}$ is the orthogonal projection onto the orthogonal complement of $\tilde{\mathbf{L}}$.

Let $\mu_{\tilde{\mathbf{L}}} = \max(\mu_{\tilde{\mathbf{U}}}, \mu_{\tilde{\mathbf{V}}})$. Simultaneously, we have

- (r1) Φ has $(1/2)\gamma_1 n_1 \leq m \leq (3/2)\gamma_1 n_1$ rows,
- (r2) each column of $\tilde{\mathbf{L}}_\Omega$ has at least $4r\mu_{\mathbf{L}} \log(2r)$ observed entries, and
- (r3) p satisfies $p \geq C_p \frac{\mu_{\tilde{\mathbf{L}}}^2 r^2 \log^3(4n_{\mathbf{L}})}{m}$.

The next result guarantees that when the column sampling parameter γ_2 is sufficiently large, exact outlier detection may be achieved. This is formalized in Lemma VI.4, and we provide the proof in Section VI-E2.

Lemma VI.4. Suppose $\tilde{\mathbf{L}}, \tilde{\mathbf{C}} \in \mathbb{R}^{m \times n_2}$ satisfy the structural conditions (g1)-(g4) with k satisfying (15), and (r1)-(r3) hold. Given $\delta \in (0, 1)$, suppose the column sampling parameter γ_2 satisfies (17) and λ satisfies (18), then the following claims hold simultaneously with probability at least $1 - \delta$:

- (I) $\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\tilde{\mathbf{C}}}$, i.e. the estimate of the outlier identities is exact, and
- (II) the total number of measurements collected is no greater than $\frac{3}{2}p\gamma_1 n_1 n_2$.

The overall results of Theorem II.3 follow by combining two intermediate lemmata via the union bound.

1) *Proof of Lemma VI.3:* Part 1: We first verify (r1)-(r3). We show that when γ_1 satisfies

$$\gamma_1 \geq \max \left\{ \frac{2r\mu_{\mathbf{U}} \log(2r)}{n_1 p}, \frac{8 \log \frac{4n_{\mathbf{L}}}{\delta}}{n_1 p}, \frac{10r\mu_{\mathbf{U}} \log \frac{4r}{\delta}}{n_1} \right\}, \quad (44)$$

then with high probability, we have

- (h1) Φ has $(1/2)\gamma_1 n_1 \leq m \leq (3/2)\gamma_1 n_1$ rows,
- (h2) each column of $\tilde{\mathbf{L}}$ has at least $4r\mu_{\mathbf{U}} \log(2r)$ observed entries,
- (h3) $\sigma_1^2(\Phi \mathbf{U}) \leq (3/2)\gamma_1$, and
- (h4) $\sigma_r^2(\Phi \mathbf{U}) \geq (1/2)\gamma_1$.

Let $\mathcal{E}_1, \dots, \mathcal{E}_4$ denote the events that (h1)-(h4) hold respectively. Then $\Pr \left(\left\{ \bigcap_{i=1}^4 \mathcal{E}_i \right\}^c \right) \leq \sum_{i=1}^4 \Pr(\mathcal{E}_i^c)$, and we consider each term in the sum as follows.

First, since m is a Binomial(n_1, γ_1) random variable, we bound its tails using [41, Theorem 2.3 (b-c)]. This gives that $\Pr(m > 3\gamma_1 n_1/2) \leq \exp(-3\gamma_1 n_1/28)$ and $\Pr(m < \gamma_1 n_1/2) \leq \exp(-\gamma_1 n_1/8)$. By union bound, we obtain that $\Pr(\mathcal{E}_1^c) \leq \exp(-3\gamma_1 n_1/28) + \exp(-\gamma_1 n_1/8)$.

Next, the number of observed entries in each column is a Binomial($n_1, p\gamma_1$). Leveraging the result in [42], we have that the minimum number of observed entries requested in the non-zero column of \mathbf{L} is $4r\mu_{\mathbf{U}} \log(2r)$ for that column to be recovered correctly with probability 1. Therefore, we need $\gamma_1 n_1 p/2 \geq 4r\mu_{\mathbf{U}} \log(2r)$, which is equivalent to

$$\gamma_1 \geq \frac{2r\mu_{\mathbf{U}} \log(2r)}{n_1 p}. \quad (45)$$

Further, by the union bound, we have

$$\begin{aligned} \Pr(\cup_{j \in \mathcal{I}_{\mathbf{L}}} \{|I_j| \leq \gamma_1 n_1 p/2\}) &\leq \sum_{j \in \mathcal{I}_{\mathbf{L}}} \Pr(|I_j| \leq \gamma_1 n_1 p/2) \\ &\leq n_{\mathbf{L}} \exp\left\{-\frac{\gamma_1 n_1 p}{8}\right\}, \end{aligned}$$

Then we have $\Pr(\mathcal{E}_2^c) \leq n_{\mathbf{L}} \exp\left\{-\frac{\gamma_1 n_1 p}{8}\right\}$.

Finally, applying [43, Corollary 5.2], we obtain

$$\begin{aligned} \Pr(\mathcal{E}_3^c) &= \Pr(\sigma_1^2(\Phi \mathbf{U}) \geq 3\gamma_1/2) \leq r \cdot (9/10)^{\frac{\gamma_1 n_1}{r\mu_{\mathbf{U}}}} \text{ and} \\ \Pr(\mathcal{E}_4^c) &= \Pr(\sigma_r^2(\Phi \mathbf{U}) \leq \gamma_1/2) \leq r \cdot (9/10)^{\frac{\gamma_1 n_1}{r\mu_{\mathbf{U}}}}. \end{aligned}$$

Putting these results together, we have

$$\begin{aligned} \Pr \left(\left\{ \bigcap_{i=1}^4 \mathcal{E}_i \right\}^c \right) &\leq \exp(-3\gamma_1 n_1/28) + \exp(-\gamma_1 n_1/8) \\ &\quad + n_{\mathbf{L}} \exp\left\{-\frac{\gamma_1 n_1 p}{8}\right\} + 2r \cdot (9/10)^{\frac{\gamma_1 n_1}{r\mu_{\mathbf{U}}}}. \end{aligned} \quad (46)$$

The R.H.S. of (46) is upper bounded by δ given that each term in the sum is no larger than $\delta/4$. This requires

$$\gamma_1 \geq \left\{ \frac{8 \log \frac{4n_{\mathbf{L}}}{\delta}}{n_1 p}, \frac{10r\mu_{\mathbf{U}} \log \frac{4r}{\delta}}{n_1} \right\}. \quad (47)$$

Combining (45) and (47), we have (44).

The condition (r1) and (r2) follow directly from (h1) and (h2). Next, we verify (r3). Given $m \geq \gamma_1 n_1/2$, for successful outlier identification via matrix completion [14], we require

$$p \geq C_p \frac{81\mu_{\mathbf{L}}^2 r^2 \log^3(4n_{\mathbf{L}})}{\gamma_1 n_1/2} \geq C_p \frac{\mu_{\tilde{\mathbf{L}}}^2 r^2 \log^3(4n_{\mathbf{L}})}{m}, \quad (48)$$

where $\mu_{\tilde{\mathbf{L}}} = 9\mu_{\mathbf{L}}$ and (r3) follows. This requires

$$\gamma_1 \geq C_p \frac{162\mu_{\mathbf{L}}^2 r^2 \log^3(4n_{\mathbf{L}})}{p n_1} = \frac{162p_l}{p}. \quad (49)$$

Combining (44) and (49), we have the bound (16) for γ_1 .

Part 2: Next, we show that $(\tilde{\mathbf{g}}1)$ – $(\tilde{\mathbf{g}}4)$ follow directly when $(\mathbf{g}1)$ – $(\mathbf{g}4)$, $(\mathbf{h}3)$ and $(\mathbf{h}4)$ hold. The condition $(\tilde{\mathbf{g}}1)$ and the first part of $(\tilde{\mathbf{g}}3)$ ($\mu_{\tilde{\mathbf{U}}} = 9\mu_{\mathbf{U}}$) follow directly from $(\mathbf{h}3)$ and $(\mathbf{h}4)$ with the argument in [15] (Lemma III.2). The second part of $(\tilde{\mathbf{g}}3)$ ($\mu_{\tilde{\mathbf{V}}} = \mu_{\mathbf{V}}$) is based on the fact that $\tilde{\mathbf{L}}$ is a row submatrix of \mathbf{L} , where $\tilde{\mathbf{L}}$ and \mathbf{L} share an identical row spaces since $\text{rank}(\tilde{\mathbf{L}}) = \text{rank}(\mathbf{L}) = r$. Therefore, they have the same column incoherence parameter (two right singular vectors are only of a difference of rotation). The condition $(\tilde{\mathbf{g}}2)$ is a direct result from $(\mathbf{g}2)$. The condition $(\tilde{\mathbf{g}}4)$ is a direct result from $(\mathbf{g}4)$, since for any row submatrix $\tilde{\mathbf{C}}_{\Omega}$ of \mathbf{C}_{Ω} , where Φ has $m \geq (1/2)\gamma_1 n_1 \geq r\mu_{\mathbf{U}} \log(2r)/p$ rows, the orthogonal projection of nonzero columns of $\Phi \mathbf{P}_{\Omega} \mathbf{C}$ onto the orthogonal complement of column space of $\tilde{\mathbf{L}}$ is not zero.

2) *Proof of Lemma VI.4:* For notational convenience, we introduce

$$\tilde{\mathbf{M}} = \tilde{\mathbf{M}}\mathbf{S}, \quad \tilde{\mathbf{L}} = \tilde{\mathbf{L}}\mathbf{S}, \quad \text{and} \quad \tilde{\mathbf{C}} = \tilde{\mathbf{C}}\mathbf{S}.$$

Using a straightforward modification of the intermediate result of Lemma III.2 in [15] again with a slightly different constant ($\delta/9$ rather than $\delta/5$), we have if γ_2 satisfies

$$\gamma_2 \geq \max \left\{ \frac{200 \log(\frac{9}{\delta})}{n_{\mathbf{L}}}, \frac{10r\mu_{\mathbf{V}} \log(\frac{9r}{\delta})}{n_{\mathbf{L}}}, \frac{200 \log(\frac{9}{\delta})}{k_u} \right\}, \quad (50)$$

then with probability at least $1 - \frac{8}{9}\delta$, we have

- (b1) \mathbf{S} has $(1/2)\gamma_2 n_2 \leq |\mathcal{S}_2| \leq (3/2)\gamma_2 n_2$ columns,
- (b2) $\tilde{\mathbf{L}}$ has $(1/2)\gamma_2 n_{\mathbf{L}} \leq n_{\tilde{\mathbf{L}}} \leq (3/2)\gamma_2 n_{\mathbf{L}}$ nonzero columns,
- (b3) $\tilde{\mathbf{C}}$ has at most $(3/2)\gamma_2 k_u$ nonzero columns,
- (b4) $\sigma_1^2(\tilde{\mathbf{V}}^* \mathbf{S}) \leq (3/2)\gamma_2$, and
- (b5) $\sigma_r^2(\tilde{\mathbf{V}}^* \mathbf{S}) \geq (1/2)\gamma_2$.

Moreover, if (b1)–(b5) hold, we have the following structural properties of $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{C}}$:

- ($\tilde{\mathbf{g}}1$) $\text{rank}(\tilde{\mathbf{L}}) = r$,
- ($\tilde{\mathbf{g}}2$) $\tilde{\mathbf{L}}$ has $n_{\tilde{\mathbf{L}}} \leq \frac{3}{2}\gamma_2 n_{\mathbf{L}}$ nonzero columns,
- ($\tilde{\mathbf{g}}3$) $\tilde{\mathbf{L}}$ satisfies the *row and column incoherence properties* with parameters $\mu_{\tilde{\mathbf{U}}} = 9\mu_{\mathbf{U}}$ and $\mu_{\tilde{\mathbf{V}}} = 9\mu_{\mathbf{V}}$ respectively, and
- ($\tilde{\mathbf{g}}4$) $\mathcal{I}_{\tilde{\mathbf{C}}} \triangleq \{j : \|\mathbf{P}_{\tilde{\mathbf{L}}^\perp} \tilde{\mathbf{C}}_{:,j}\|_2 > 0\}$ with $|\mathcal{I}_{\tilde{\mathbf{C}}}| = \tilde{k}$, where $\tilde{\mathcal{L}}$ denotes the subspace spanned by columns of $\tilde{\mathbf{L}}$ and $\mathbf{P}_{\tilde{\mathbf{L}}^\perp}$ is the orthogonal projection operation onto the orthogonal complement of $\tilde{\mathcal{L}}$.

Next, we show that MP in Step 1 of RACOS-I succeeds with high probability under proper conditions. Let $\mu_{\tilde{\mathbf{L}}} = \max(\mu_{\tilde{\mathbf{U}}}, \mu_{\tilde{\mathbf{V}}})$, \tilde{k} be the number of outliers in $\tilde{\mathbf{M}}$, and $\tilde{n}_2 = |\mathcal{S}_2|$ be the number columns of $\tilde{\mathbf{M}}$. We formalize the result in Lemma VI.5 without proof.

Lemma VI.5 (Adapted from Theorem 1 in [14]). *Suppose $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{C}}$ satisfy the structural conditions ($\tilde{\mathbf{g}}1$)–($\tilde{\mathbf{g}}4$) with $n_{\tilde{\mathbf{L}}} \geq m \geq 32$. If (r, k, p) satisfies (48) and*

$$\tilde{k} \leq \frac{p^2 \tilde{n}_2}{p^2 + C_{\tilde{k}}(1 + \frac{\mu_{\tilde{\mathbf{L}}} r}{p\sqrt{m}})\mu_{\tilde{\mathbf{L}}}^3 r^3 \log^6(4n_{\tilde{\mathbf{L}}})}, \quad (51)$$

, where $C_{\tilde{k}}$ is a constant, and λ satisfies

$$\lambda = \frac{1}{48} \sqrt{\frac{p}{kr\mu_{\tilde{\mathbf{L}}} \log^2(4n_{\tilde{\mathbf{L}}})}},$$

then MP returns $\{\hat{\mathbf{L}}_{(1)}, \hat{\mathbf{C}}_{(1)}\}$ such that $\hat{\mathbf{L}}_{(1)}$ has the same column space with $\tilde{\mathbf{L}}$, and $\hat{\mathbf{C}}_{(1)}$ has the same column support with $\tilde{\mathbf{C}}$, with probability at least $1 - \tilde{C}_{\gamma_2} \tilde{n}_2^{-5}$ for some positive constant \tilde{C}_{γ_2} .

We start by verifying that p and \tilde{k} satisfy the required bound for the data matrix $\tilde{\mathbf{M}}$ in Lemma VI.5. Given p satisfying the bound in (r3) of Lemma VI.3 and $n_{\mathbf{L}} \geq n_{\tilde{\mathbf{L}}}$, we have that

$$p \geq C_p \frac{\mu_{\tilde{\mathbf{L}}}^2 r^2 \log^3(4n_{\tilde{\mathbf{L}}})}{m}, \quad (52)$$

where $\mu_{\tilde{\mathbf{L}}} = \mu_{\tilde{\mathbf{L}}} = 9\mu_{\mathbf{L}}$.

For \tilde{k} , when (50) and $\frac{200}{\gamma_2} \log(\frac{9}{\delta}) \leq k \leq k_u$ hold, with probability at least $1 - \delta/9$, we have

$$\begin{aligned} \tilde{k} &\leq \frac{3}{2}\gamma_2 k_u = \frac{p^2 \gamma_2 n_2 / 2}{p^2 + C_k(1 + \frac{3\sqrt{6}\mu_{\mathbf{L}} r}{p\sqrt{n_1}})n_{\mathbf{L}}^3 r^3 \log^6(4n_{\mathbf{L}})} \\ &\leq \frac{p^2 \gamma_2 n_2 / 2}{p^2 + C_{\tilde{k}}(1 + \frac{9\mu_{\mathbf{L}} r}{p\sqrt{\frac{3}{2}\gamma_1 n_1}})(9\mu_{\mathbf{L}})^3 r^3 \log^6(4n_{\mathbf{L}})} \\ &\leq \frac{p^2 \tilde{n}_2}{p^2 + C_{\tilde{k}}(1 + \frac{\mu_{\tilde{\mathbf{L}}} r}{p\sqrt{m}})\mu_{\tilde{\mathbf{L}}}^3 r^3 \log^6(4n_{\tilde{\mathbf{L}}})}, \end{aligned}$$

where $C_{\tilde{k}} = C_k/729$. When $k < \frac{200}{\gamma_2} \log(\frac{9}{\delta})$, we have from Lemma 4.1 in [16] that $\Pr(\tilde{k}_2 \leq t) \geq \Pr(\tilde{k}_1 \leq t)$ for any $t \in [0, \infty)$, where \tilde{k}_1 and \tilde{k}_2 are Hypergeometric random variables with distributions $\text{Hyp}(n_2, \tilde{n}_2, k_1)$ and $\text{Hyp}(n_2, \tilde{n}_2, k_2)$ respectively. This implies with probability at least $1 - 9/\delta$, we have

$$\tilde{k} \leq \frac{3}{2}\gamma_2 k_u \leq \frac{p^2 \tilde{n}_2}{p^2 + C_{\tilde{k}}(1 + \frac{\mu_{\tilde{\mathbf{L}}} r}{p\sqrt{m}})\mu_{\tilde{\mathbf{L}}}^3 r^3 \log^6(4n_{\tilde{\mathbf{L}}})}. \quad (53)$$

This also verifies (b3).

From $n_{\mathbf{L}} \geq \frac{\delta e^{8p}}{4}$ and $\gamma_1 \geq \frac{8 \log \frac{4n_{\mathbf{L}}}{\delta}}{n_1 p}$, if

$$\gamma_2 \geq \frac{3\gamma_1 n_1}{n_{\mathbf{L}}}, \quad (54)$$

we have $n_{\tilde{\mathbf{L}}} \geq m \geq 32$. Exact outlier identification is achievable with high probability when (52), (53), and (51) hold.

Let $\tilde{C}_{\gamma_2} \tilde{n}_2^{-5} \leq \tilde{C}_{\gamma_2} (\frac{3}{2}\gamma_2 n_2)^{-5} \leq \delta/9$, then we have that

$$\gamma_2 \geq \frac{3}{2n_2} \left(\frac{9\tilde{C}_{\gamma_2}}{\delta} \right)^{\frac{1}{5}} = \frac{C_{\gamma_2} (\frac{1}{\delta})^{\frac{1}{5}}}{n_2}, \quad (55)$$

where $C_{\gamma_2} = \frac{3}{2}(9\tilde{C}_{\gamma_2})^{\frac{1}{5}}$. Combining (50), (54) and (55), we obtain the bound (17) for γ_2 .

Note that $\hat{\mathcal{L}}_{(1)} = \tilde{\mathcal{L}} = \hat{\mathcal{L}}$ and the number of observed entries for each non-zero column of $\tilde{\mathbf{L}}$ is at least $4r\mu_{\mathbf{U}} \log(2r)$ from (r2) of Lemma VI.3. Then we have that for any $j \in \mathcal{I}_{\tilde{\mathbf{L}}}$, $\|\mathbf{P}_{\tilde{\mathbf{L}}_{\tilde{\mathcal{L}}_j}^\perp} \tilde{\mathbf{L}}_{\mathcal{I}_j, j}\| = 0$ with probability 1 from [42] (Theorem 1), and for any $j \in \mathcal{I}_{\tilde{\mathbf{C}}}$, $\|\mathbf{P}_{\tilde{\mathbf{L}}^\perp} \tilde{\mathbf{C}}_{\mathcal{I}_j, j}\| > 0$ from ($\tilde{\mathbf{g}}4$). Therefore, we have $\hat{\mathcal{I}}_{\tilde{\mathbf{C}}} = \mathcal{I}_{\tilde{\mathbf{C}}}$, thus Claim (I) follows. Claim (II) holds directly from (r1). Finally, the overall result holds with probability at least $1 - \delta$ using the union bound.

F. Proof of Theorem II.4

The analysis of Theorem II.4 is analogous to that of Theorem II.3. For completeness, we provide the intermediate results here. We first show that the structural conditions for $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{C}}$ provided that the row sampling parameter γ_1 is sufficiently large. This is formalized in Lemma VI.6, and we provide the proof in Section VI-F1.

Lemma VI.6. Suppose $\mathbf{L}, \mathbf{C} \in \mathbb{R}^{n_1 \times n_2}$ satisfy the structural conditions (g1)-(g4) with p satisfying (19). Given $\delta \in (0, 1)$, suppose $\Phi \in \mathbb{R}^{m \times n_1}$ is a row sampling matrix with the sampling parameter γ_1 satisfying (16). Then, with probability at least $1 - \delta$, the components $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{C}}$ satisfy

- (g1) $\text{rank}(\tilde{\mathbf{L}}) = r$,
- (g2) $\tilde{\mathbf{L}}$ has $n_{\mathbf{L}}$ nonzero columns,
- (g3) $\tilde{\mathbf{L}}$ satisfies the row and column incoherence property with parameters $\mu_{\tilde{\mathbf{U}}} = 9\mu_{\mathbf{U}}$ and $\mu_{\tilde{\mathbf{V}}} = \mu_{\mathbf{V}}$ respectively, and
- (g4) $\mathcal{I}_{\tilde{\mathbf{C}}} \triangleq \{j \in [n_2] : \|(\mathbf{P}_{\tilde{\mathbf{L}}^\perp}(\tilde{\mathbf{C}}\Omega))_{:,j}\|_2 > 0\} = \mathcal{I}_{\mathbf{C}}$, where $\tilde{\mathbf{L}}$ denotes the subspace spanned by columns of $\tilde{\mathbf{L}}$, and $\mathbf{P}_{\tilde{\mathbf{L}}^\perp}$ is the orthogonal projection onto the orthogonal complement of $\tilde{\mathbf{L}}$ in \mathbb{R}^m .

Let $\mu_{\tilde{\mathbf{L}}} = \max(\mu_{\tilde{\mathbf{U}}}, \mu_{\tilde{\mathbf{V}}})$. Simultaneously, we have

- (r1) Φ has $(1/2)\gamma_1 n_1 \leq m \leq (3/2)\gamma_1 n_1$ rows,
- (r2) each column of $\tilde{\mathbf{L}}\Omega$ has at least $4r\mu_{\mathbf{L}} \log(2r)$ observed entries, and
- (r3) p satisfies $p \geq C_p \frac{\mu_{\tilde{\mathbf{L}}}^2 r^2 \log^3(4n_{\mathbf{L}})}{m}$.

The next result guarantees that when the column sampling parameter γ_2 is sufficiently large, exact outlier detection may be achieved. This is formalized in Lemma VI.7, and we provide the proof in Section VI-F2.

Lemma VI.7. Suppose $\tilde{\mathbf{L}}, \tilde{\mathbf{C}} \in \mathbb{R}^{m \times n_2}$ satisfy the conditions (g1)-(g4) with k satisfying (20), and (r1)-(r3) hold. Given $\delta \in (0, 1)$, suppose the column sampling parameter γ_2 satisfies (17) and λ satisfies (21), then the following hold simultaneously with probability at least $1 - \delta$:

- (I) $\hat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\tilde{\mathbf{C}}}$, i.e. the estimate of the outlier identities is exact, and
- (II) the total number of measurements collected is no greater than $\frac{3}{2}p\gamma_1 n_1 n_2$.

The overall result of Theorem II.4 follows by combining two intermediate results via the union bound.

1) *Proof of Lemma VI.6:* The analysis follows directly from that of Lemma VI.3, except for (r3). Given $m \geq \gamma_1 n_1 / 2$, for successful outlier identification via MP [14], we need p to satisfy

$$\begin{aligned} p &\geq C_p \left(1 + \frac{1}{\varphi}\right) \frac{9\mu_{\mathbf{L}} r \log^2(2n_2)}{\gamma_1 n_1 / 2} \\ &\geq C_p \left(1 + \frac{1}{\varphi}\right) \frac{\mu_{\tilde{\mathbf{L}}} r \log^2(2\tilde{n}_2)}{m}, \end{aligned} \quad (56)$$

which is implied by (19), where $\mu_{\tilde{\mathbf{L}}} = 9\mu_{\mathbf{L}}$ and (r3) follows. This requires

$$\gamma_1 \geq C_p \left(1 + \frac{1}{\varphi}\right) \frac{18\mu_{\mathbf{L}} r \log^2(2n_2)}{pn_1} = \frac{18p_l}{p}. \quad (57)$$

Combining (44) and (56), we have the bound (16) for γ_1 .

2) *Proof of Lemma VI.7:* The analysis follows directly from that of Lemma VI.4, except the bound of k . For successful outlier identification via MP [14], we need \tilde{k} to satisfy

$$\tilde{k} \leq C_{\tilde{k}} \frac{\varphi}{1 + \varphi\sqrt{\varphi}} \frac{pn_{\tilde{\mathbf{L}}}}{\mu_{\tilde{\mathbf{L}}}^{3/2} r^{3/2} \log^3(2n_2)},$$

where $C_{\tilde{k}}$ is a constant, which holds since $\tilde{k} \leq \frac{3}{2}\gamma_2 k_u$ and k_u satisfies (20).

G. Proof of Lemma VI.1

We leverage the intermediate result of the proof of Theorem 2 in [9], which provides the estimation error bounds of both low-rank and outlier components w.r.t. the noise in term of the Frobenius norm. However, we are interested in the the estimation error bound of the low-rank component in terms of the spectral norm, which are the main technical differences in our proof here.

To start with, we define two operators that return the subgradient of $\|\mathbf{L}\|_*$ and $\|\mathbf{C}\|_{1,2}$ in the following lemma. Recall that $\mathbf{P}_{\mathcal{L}}(\cdot)$ is an orthogonal projection operator that project a matrix to the column space \mathcal{L} of \mathbf{L} , and $\mathbf{P}_{\mathcal{I}_{\mathbf{C}}}(\cdot)$ is a projection operator that leave columns in the support set $\mathcal{I}_{\mathbf{C}}$ unchanged and set the other columns to be 0.

Definition VI.1. Let $\mathbf{M} = \mathbf{L}' + \mathbf{C}'$, where $\mathbf{P}_{\mathcal{L}}(\mathbf{L}') = \mathbf{L}'$ and $\mathbf{P}_{\mathcal{I}_{\mathbf{C}}}(\mathbf{C}') = \mathbf{C}'$. Given the compact SVD of \mathbf{L}' as $\mathbf{L}' = \mathbf{U}'\Sigma'\mathbf{V}'^T$ and the column support of \mathbf{C}' as $\mathcal{I}'_{\mathbf{C}}$, we define the following:

$$\begin{aligned} \mathfrak{R}(\mathbf{L}') &\triangleq \mathbf{U}'\mathbf{V}'^T; \\ \mathfrak{G}(\mathbf{C}') &\triangleq \{\mathbf{H} \in \mathbb{R}_{n_1 \times n_2} | \mathbf{P}_{\mathcal{I}_{\mathbf{C}}}(\mathbf{H}) = \mathbf{0}; \forall i \in \mathcal{I}'_{\mathbf{C}} \subseteq \mathcal{I}_{\mathbf{C}}, \mathbf{H}_{:,i} = \mathbf{C}'_{:,i} / \|\mathbf{C}'_{:,i}\|_2; \forall i \in \mathcal{I}_{\mathbf{C}} \cap (\mathcal{I}'_{\mathbf{C}})^c, \|\mathbf{H}_{:,i}\|_2 < 1\}. \end{aligned}$$

Consider the noisy OP problem (29). Let $\tilde{\mathbf{M}}_0 = \tilde{\mathbf{L}}_1 + \tilde{\mathbf{C}}_1$, where $\mathbf{P}_{\tilde{\mathcal{L}}}(\tilde{\mathbf{L}}_1) = \tilde{\mathbf{L}}_1$ and $\mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\tilde{\mathbf{C}}_1) = \tilde{\mathbf{C}}_1$. For $\tilde{\mathbf{L}}_1 = \tilde{\mathbf{U}}_1 \tilde{\Sigma}_1 \tilde{\mathbf{V}}_1^T$ and $\tilde{\mathbf{L}} = \tilde{\mathbf{U}} \tilde{\Sigma} \tilde{\mathbf{V}}^T$, there exists an orthonormal matrix $\tilde{\mathbf{V}} \in \mathbb{R}^{r \times \tilde{n}_2}$, such that $\tilde{\mathbf{U}}_1 \tilde{\mathbf{V}}_1^T = \tilde{\mathbf{U}} \tilde{\mathbf{V}}^T$. Further let $\mathbf{P}_{\mathcal{T}(\tilde{\mathbf{L}}_1)}$ be the projection onto the space spanned by $\tilde{\mathbf{U}}_1$ and $\tilde{\mathbf{V}}_1$, which is given by $\mathbf{P}_{\mathcal{T}(\tilde{\mathbf{L}}_1)} = \mathbf{P}_{\tilde{\mathbf{U}}_1} + \mathbf{P}_{\tilde{\mathbf{V}}_1} - \mathbf{P}_{\tilde{\mathbf{U}}_1} \mathbf{P}_{\tilde{\mathbf{V}}_1}$, $\tilde{\mathbf{N}}_{\mathbf{L}} = \tilde{\mathbf{L}} - \tilde{\mathbf{L}}_1$, and $\tilde{\mathbf{N}}_{\mathbf{C}} = \tilde{\mathbf{C}} - \tilde{\mathbf{C}}_1$, thus $\tilde{\mathbf{N}} = \tilde{\mathbf{N}}_{\mathbf{L}} + \tilde{\mathbf{N}}_{\mathbf{C}}$. Define $\tilde{\mathbf{N}}_{\mathbf{L}}^+ = \tilde{\mathbf{N}}_{\mathbf{L}} - \mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}} \mathbf{P}_{\tilde{\mathcal{L}}}(\tilde{\mathbf{N}}_{\mathbf{L}})$, $\tilde{\mathbf{N}}_{\mathbf{C}}^+ = \tilde{\mathbf{N}}_{\mathbf{C}} - \mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}} \mathbf{P}_{\tilde{\mathcal{L}}}(\tilde{\mathbf{N}}_{\mathbf{C}})$, and $\tilde{\mathbf{N}}^+ = \tilde{\mathbf{N}} - \mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}} \mathbf{P}_{\tilde{\mathcal{L}}}(\tilde{\mathbf{N}})$. It is shown in [9] (Lemma 11) that for any $\mathbf{X} \in \mathbb{R}^{m \times \tilde{n}_2}$

$$\mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}} \mathbf{P}_{\tilde{\mathbf{V}}} \mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\mathbf{X}) = \mathbf{X} (\mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\tilde{\mathbf{V}}^T))^T \mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\tilde{\mathbf{V}}^T),$$

and correspondingly for some ψ , we have

$$\begin{aligned} \|\mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}} \mathbf{P}_{\tilde{\mathbf{V}}} \mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\mathbf{X})\|_2 &= \|\mathbf{X} (\mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\tilde{\mathbf{V}}^T))^T \mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\tilde{\mathbf{V}}^T)\|_2 \\ &\leq \|\mathbf{X}\|_2 \|(\mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\tilde{\mathbf{V}}^T))^T \mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\tilde{\mathbf{V}}^T)\|_2 \leq \psi \|\mathbf{X}\|_2, \end{aligned} \quad (58)$$

where the last inequality follows from the bound $\|(\mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\tilde{\mathbf{V}}^T))^T \mathbf{P}_{\mathcal{I}_{\tilde{\mathbf{C}}}}(\tilde{\mathbf{V}}^T)\|_2 \leq \psi$. It is shown in [9] that if $\tilde{k} = |\mathcal{I}_{\tilde{\mathbf{C}}}|$ satisfies (26), then we have $\psi < \frac{1}{4}$.

The main body of the proof is to construct a dual certificate to guarantee that the optimal solution pair of (29) is “close” to a pair of $(\tilde{\mathbf{L}}_0, \tilde{\mathbf{C}}_0)$, which has the correct column space and the correct column support respectively, in terms of the spectral

norm, given the $\ell_{1,2}$ -norm of the noise term. We demonstrate this in Lemma VI.8.

Lemma VI.8 (Adapted from Theorem 5 in [9]). *Let $\hat{\mathbf{L}}, \hat{\mathbf{C}}$ be an optimal solution pair of (29). Suppose $\lambda = \frac{\sqrt{9+1024\mu_L r}}{14\sqrt{n_2}} < 1$ and $\psi < \frac{1}{4}$. Let $\check{\mathbf{M}}_0 = \check{\mathbf{L}}_1 + \check{\mathbf{C}}_1$, where $\mathbf{P}_{\check{\mathcal{L}}}(\check{\mathbf{L}}_1) = \check{\mathbf{L}}_1$ and $\mathbf{P}_{\check{\mathcal{C}}}(\check{\mathbf{C}}_1) = \check{\mathbf{C}}_1$. If there exists \mathbf{Q} such that*

$$\begin{aligned} \mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)}(\mathbf{Q}) &= \Re(\check{\mathbf{L}}_1), \quad \|\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\mathbf{Q})\|_2 \leq 1/2, \\ \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\mathbf{Q})/\lambda &\in \mathfrak{G}(\check{\mathbf{C}}_1), \quad \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\mathbf{Q})\|_{\infty,2} \leq \lambda/2, \end{aligned}$$

then there exists a pair $(\check{\mathbf{L}}_0, \check{\mathbf{C}}_0)$ such that $\check{\mathbf{M}}_0 = \check{\mathbf{L}}_0 + \check{\mathbf{C}}_0$, $\mathbf{P}_{\check{\mathcal{L}}}(\check{\mathbf{L}}_0) = \check{\mathbf{L}}_0$ and $\mathbf{P}_{\check{\mathcal{C}}}(\check{\mathbf{C}}_0) = \check{\mathbf{C}}_0$, and

$$\|\hat{\mathbf{L}} - \check{\mathbf{L}}_0\|_2 \leq 10\|\check{\mathbf{N}}\|_{1,2}, \quad \|\hat{\mathbf{C}} - \check{\mathbf{C}}_0\|_2 \leq 9\|\check{\mathbf{N}}\|_{1,2}.$$

Proof of Lemma VI.8. Let us introduce two quantities \mathbf{W} and \mathbf{F} that are related to the subgradient of $\|\hat{\mathbf{L}}_1\|_*$ and $\|\hat{\mathbf{C}}_1\|_{1,2}$. We have from Theorem 3 in [9] that for any fixed perturbation $\Delta \neq 0$, $(\hat{\mathbf{L}}_1 + \Delta, \hat{\mathbf{C}}_1 - \Delta)$ is strictly worse than $(\hat{\mathbf{L}}_1, \hat{\mathbf{C}}_1)$, unless $\Delta \in \mathbf{P}_{\check{\mathcal{L}}} \cap \mathbf{P}_{\check{\mathcal{C}}}$. Let \mathbf{W} be such that $\|\mathbf{W}\|_2 = 1$, $\langle \mathbf{W}, \mathbf{P}_{\mathcal{T}(\hat{\mathbf{L}})^\perp}(\Delta) \rangle = \|\mathbf{P}_{\mathcal{T}(\hat{\mathbf{L}})^\perp}(\Delta)\|_*$, and $\mathbf{P}_{\mathcal{T}(\hat{\mathbf{L}})}(\mathbf{W}) = 0$. Let \mathbf{F} be such that

$$\mathbf{F}_{:,i} = \begin{cases} \frac{-\Delta_{:,i}}{\|\Delta_{:,i}\|_2}, & \text{if } i \notin \mathcal{I}_{\check{\mathcal{C}}}, \text{ and } \Delta_{:,i} \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Then $\mathbf{P}_{\mathcal{T}(\hat{\mathbf{L}})}(\mathbf{Q}) + \mathbf{W}$ is a subgradient of $\|\hat{\mathbf{L}}_1\|_*$ and $\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\mathbf{Q})/\lambda + \mathbf{F}$ is a subgradient of $\|\hat{\mathbf{C}}_1\|_{1,2}$.

From the optimality of $\hat{\mathbf{L}}$ and $\hat{\mathbf{C}}$, we have

$$\begin{aligned} \|\check{\mathbf{L}}_1\|_* + \lambda\|\check{\mathbf{C}}_1\|_{1,2} &\geq \|\hat{\mathbf{L}}_1\|_* + \lambda\|\hat{\mathbf{C}}_1\|_{1,2} \\ &\geq \|\check{\mathbf{L}}_1\|_* + \lambda\|\check{\mathbf{C}}_1\|_{1,2} + \langle \mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)}(\mathbf{Q}) + \mathbf{W}, \check{\mathbf{N}}_{\mathbf{L}} \rangle \\ &\quad + \lambda\langle \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\mathbf{Q})/\lambda + \mathbf{F}, \check{\mathbf{N}}_{\mathbf{C}} \rangle \\ &\stackrel{(i)}{=} \|\check{\mathbf{L}}_1\|_* + \lambda\|\check{\mathbf{C}}_1\|_{1,2} + \|\mathbf{P}_{\mathcal{T}(\hat{\mathbf{L}})^\perp}(\check{\mathbf{N}}_{\mathbf{L}})\|_* + \lambda\|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}})\|_{1,2} \\ &\quad + \langle \mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)}(\mathbf{Q}), \check{\mathbf{N}}_{\mathbf{L}} \rangle + \langle \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\mathbf{Q}), \check{\mathbf{N}}_{\mathbf{C}} \rangle \\ &= \|\check{\mathbf{L}}_1\|_* + \lambda\|\check{\mathbf{C}}_1\|_{1,2} + \|\mathbf{P}_{\mathcal{T}(\hat{\mathbf{L}})^\perp}(\check{\mathbf{N}}_{\mathbf{L}})\|_* + \lambda\|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}})\|_{1,2} \\ &\quad - \langle \mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\mathbf{Q}), \check{\mathbf{N}}_{\mathbf{L}} \rangle - \langle \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\mathbf{Q}), \check{\mathbf{N}}_{\mathbf{C}} \rangle + \langle \mathbf{Q}, \check{\mathbf{N}}_{\mathbf{L}} + \check{\mathbf{N}}_{\mathbf{C}} \rangle \\ &\geq \|\check{\mathbf{L}}_1\|_* + \lambda\|\check{\mathbf{C}}_1\|_{1,2} + (1 - \|\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\mathbf{Q})\|)\|\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\check{\mathbf{N}}_{\mathbf{L}})\|_* \\ &\quad + (\lambda - \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\mathbf{Q})\|_{\infty,2})\|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}})\|_{1,2} + \langle \mathbf{Q}, \check{\mathbf{N}} \rangle \\ &\geq \|\check{\mathbf{L}}_1\|_* + \lambda\|\check{\mathbf{C}}_1\|_{1,2} + \frac{1}{2}\|\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\check{\mathbf{N}}_{\mathbf{L}})\|_* \\ &\quad + \frac{\lambda}{2}\|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}})\|_{1,2} - \|\check{\mathbf{N}}\|_{1,2}\|\mathbf{Q}\|_{\infty,2}, \end{aligned} \quad (59)$$

where (i) is from the choice of \mathbf{W} and \mathbf{F} above. From (59), we have

$$\begin{aligned} \|\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\check{\mathbf{N}}_{\mathbf{L}})\|_2 &\leq \|\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\check{\mathbf{N}}_{\mathbf{L}})\|_* \\ &\leq 2\lambda\|\check{\mathbf{N}}\|_{1,2}\|\mathbf{Q}\|_{\infty,2} \leq 2\lambda\|\check{\mathbf{N}}\|_{1,2}, \quad (60) \\ \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}})\|_2 &\leq \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}})\|_{1,2} \\ &\leq 2\|\check{\mathbf{N}}\|_{1,2}\|\mathbf{Q}\|_{\infty,2} \leq 2\|\check{\mathbf{N}}\|_{1,2}. \quad (61) \end{aligned}$$

From the result in [9] (eqn. (16)), we have

$$\begin{aligned} \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}^+) &= \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}) - \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\check{\mathbf{N}}_{\mathbf{L}}) - \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)}(\check{\mathbf{N}}) \\ &\quad + \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)}\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}) + \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\check{\mathbf{V}}}\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}^+). \end{aligned}$$

By triangle inequality, the equality above implies

$$\begin{aligned} \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}^+)\|_2 &\leq \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}) - \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\check{\mathbf{N}}_{\mathbf{L}})\|_2 + \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\check{\mathbf{N}}_{\mathbf{L}})\|_2 \\ &\quad + \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)}\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}})\|_2 + \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\check{\mathbf{V}}}\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}^+)\|_2 \\ &\stackrel{(i)}{\leq} \|\check{\mathbf{N}}\|_2 + \|\mathbf{P}_{\mathcal{T}(\check{\mathbf{L}}_1)^\perp}(\check{\mathbf{N}}_{\mathbf{L}})\|_2 + \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}})\|_2 \\ &\quad + \psi\|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}^+)\|_2 \\ &\stackrel{(ii)}{\leq} (1 + 2\lambda + 2)\|\check{\mathbf{N}}\|_{1,2} + \psi\|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}^+)\|_2, \end{aligned} \quad (62)$$

where (i) is from (58), and (ii) is from (60), (61), and the fact $\|\check{\mathbf{N}}\|_2 \leq \|\check{\mathbf{N}}\|_F \leq \|\check{\mathbf{N}}\|_{1,2}$. From (62), we have

$$\|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}^+)\|_2 \leq \frac{(1 + 2\lambda\sqrt{n_2} + 2\sqrt{n_2})\|\check{\mathbf{N}}\|_{1,2}}{1 - \psi}. \quad (63)$$

Combining (61), (63), and the fact that $\lambda < 1$ and $\psi < \frac{1}{4}$, we have

$$\begin{aligned} \|\check{\mathbf{N}}_{\mathbf{C}}^+\|_2 &= \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}) + \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}^+)\|_{1,2} \\ &\leq \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}})\|_2 + \|\mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}(\check{\mathbf{N}}_{\mathbf{C}}^+)\|_2 \leq 9\|\check{\mathbf{N}}\|_{1,2}. \end{aligned}$$

Finally, note that we have

$$\check{\mathbf{N}}_{\mathbf{C}}^+ = (\mathbf{I} - \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\mathcal{I}_{\check{\mathcal{L}}}})(\hat{\mathbf{C}} - \check{\mathbf{C}}_1) = \hat{\mathbf{C}} - \check{\mathbf{C}}_0,$$

where $\check{\mathbf{C}}_0 = \check{\mathbf{C}}_1 + \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\mathcal{I}_{\check{\mathcal{L}}}}(\hat{\mathbf{C}} - \check{\mathbf{C}}_1)$. Since $\check{\mathbf{C}}_1 \in \mathcal{I}_{\check{\mathcal{C}}}$, this implies $\check{\mathbf{C}}_0 \in \mathcal{I}_{\check{\mathcal{C}}}$ and

$$\|\hat{\mathbf{C}} - \check{\mathbf{C}}_0\|_2 \leq 9\|\check{\mathbf{N}}\|_{1,2}.$$

Further let $\check{\mathbf{L}}_0 = \check{\mathbf{L}}_1 - \mathbf{P}_{\mathcal{I}_{\check{\mathcal{C}}}}\mathbf{P}_{\mathcal{I}_{\check{\mathcal{L}}}}(\hat{\mathbf{C}} - \check{\mathbf{C}}_1)$, we have that $\check{\mathbf{L}}_0$ and $\check{\mathbf{C}}_0$ are a pair of successful decomposition, and

$$\|\check{\mathbf{L}}_0 - \hat{\mathbf{L}}\|_2 \leq \|\check{\mathbf{N}}\|_{1,2} + \|\hat{\mathbf{C}} - \check{\mathbf{C}}_0\|_2 \leq 10\|\check{\mathbf{N}}\|_{1,2}.$$

□

H. Further Intermediate Results

In this section, we provide the statement of several intermediate results adopted in our analysis.

The first intermediate result provides the additive perturbation bound for the singular values.

Lemma VI.9 (Theorem 1 of [44]). *Suppose $n_1 \leq n_2$. Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a matrix with singular values $\{\sigma_i\}_{i=1}^{n_1}$, and $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ be a perturbation of \mathbf{A} with singular values $\{\tilde{\sigma}_i\}_{i=1}^{n_1}$. Then for any $i \in [n_1]$, the following bound holds:*

$$|\tilde{\sigma}_i - \sigma_i| \leq \|\mathbf{E}\|_2.$$

The second intermediate result provides the additive perturbation bound for the orthogonal projection.

Lemma VI.10 (Adapted from Theorem 2.2 of [45]). *Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ be a rank- r matrix with SVD*

$$\mathbf{A} = [\mathbf{U}_1 \mathbf{U}_2] \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix},$$

and $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ be a perturbation of \mathbf{A} with $\text{rank}(\tilde{\mathbf{A}}) = \text{rank}(\mathbf{A}) = r$. Let $\tilde{\mathcal{A}}$ and \mathcal{A} be the column spaces of $\tilde{\mathbf{A}}$ and \mathbf{A} respectively. Then the following bound holds:

$$\|\mathbf{P}_{\tilde{\mathcal{A}}} - \mathbf{P}_{\mathcal{A}}\|_2 \leq \min\{1, \|\mathbf{A}^\dagger\|_2 \|\mathbf{E}\mathbf{V}_1\|_2, \|\tilde{\mathbf{A}}^\dagger\|_2 \|\mathbf{U}_2^T \mathbf{E}\|_2\},$$

where \mathbf{A}^\dagger is the Moore-Penrose inverse of \mathbf{A} .

REFERENCES

- [1] B. Mehta and W. Nejdl, "Attack resistant collaborative filtering," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 75–82.
- [2] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM Computer Communication Review*. ACM, 2004, vol. 34, pp. 219–230.
- [3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [4] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NIPS*, 2006, pp. 545–552.
- [5] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," in *Proc. CVPR*, 2007.
- [6] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. CVPR*, 2012, pp. 853–860.
- [7] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optimization*, vol. 21, no. 2, pp. 572–596, 2011.
- [8] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.
- [9] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," *IEEE Trans. Inform. Theory*, vol. 58, no. 5, pp. 3047–3064, 2012.
- [10] M. McCoy and J. A. Tropp, "Two proposals for robust PCA using semidefinite programming," *Electronic J. of Statistics*, vol. 5, pp. 1123–1160, 2011.
- [11] M. Soltanolkotabi and E. Candes, "A geometric analysis of subspace clustering with outliers," *The Annals of Statistics*, vol. 40, no. 4, pp. 2195–2238, 2012.
- [12] M. Hardt and A. Moitra, "Algorithms and hardness for robust subspace recovery," in *Conf. on Learning Theory*, 2013, pp. 354–375.
- [13] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang, "Robust computation of linear models by convex relaxation," *Foundations of Computational Mathematics*, pp. 1–48, 2014.
- [14] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi, "Matrix completion with column manipulation: Near-optimal sample-robustness-rank trade-offs," *IEEE Trans. Inform. Theory*, vol. 62, no. 1, pp. 503–526, 2016.
- [15] X. Li and J. Haupt, "Identifying outliers in large matrices via randomized adaptive compressive sampling," *Trans. Signal Processing*, vol. 63, no. 7, pp. 1792–1807, 2015.
- [16] X. Li and J. Haupt, "A refined analysis for the sample complexity of adaptive compressive outlier sensing," in *IEEE Workshop on Statistical Signal Processing*, 2016.
- [17] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal recovery from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.
- [18] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [19] E. Bashan, R. Raich, and A. O. Hero, "Optimal two-stage search for sparse targets using convex criteria," *IEEE Trans Signal Processing*, vol. 56, no. 11, pp. 5389–5402, 2008.
- [20] J. Haupt, R. M. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Trans. Information Theory*, vol. 57, no. 9, pp. 6222–6235, 2011.
- [21] E. Bashan, G. Newstadt, and A. O. Hero, "Two-stage multiscale search for sparse targets," *IEEE Trans Signal Processing*, vol. 59, no. 5, pp. 2331–2341, 2011.
- [22] M. L. Malloy and R. Nowak, "Sequential testing for sparse recovery," *arXiv preprint:1212.1801*, 2012.
- [23] R. M. Castro, "Adaptive sensing performance lower bounds for sparse signal detection and support estimation," *arXiv preprint:1206.0648*, 2012.
- [24] A. Krishnamurthy, J. Sharpnack, and A. Singh, "Recovering graph-structured activations using adaptive compressive measurements," *arXiv preprint:1305.0213*, 2013.
- [25] D. Wei and A. O. Hero, "Multistage adaptive estimation of sparse signals," *IEEE J. of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 783–796, 2013.
- [26] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi, "Robust matrix completion with corrupted columns," *arXiv preprint:1102.2254*, 2011.
- [27] P. Huber, *Robust statistics*, Springer, 2011.
- [28] H. Nyquist, "Least orthogonal absolute deviations," *Computational Statistics & Data Analysis*, vol. 6, no. 4, pp. 361–367, 1988.
- [29] C. Yang, D. Robinson, and R. Vidal, "Sparse subspace clustering with missing entries," in *Proceedings of The 32nd International Conference on Machine Learning*, 2015, pp. 2463–2472.
- [30] M. Rahmani and G. Atia, "Randomized robust subspace recovery for high dimensional data matrices," *arXiv preprint arXiv:1505.05901*, 2015.
- [31] Y. She, S. Li, and D. Wu, "Robust orthogonal complement principal component analysis," *arXiv preprint arXiv:1410.1173*, 2014.
- [32] T. Zhang and G. Lerman, "A novel m-estimator for robust pca," *J. Machine Learning Research*, vol. 15, no. 1, pp. 749–808, 2014.
- [33] O. Klopp, K. Lounici, and A. Tsybakov, "Robust matrix completion," *arXiv preprint arXiv:1412.8132*, 2014.
- [34] G. Obozinski, M. J. Wainwright, and M. Jordan, "Support union recovery in high-dimensional multivariate regression," *The Annals of Statistics*, pp. 1–47, 2011.
- [35] X. Li and J. Haupt, "Locating salient group-structured image features via adaptive compressive sensing," in *GlobalSIP*, 2015.
- [36] X. Li and J. Haupt, "Outlier identification via randomized adaptive compressive sampling," in *ICASSP*, 2015.
- [37] D. Woodruff, "Sketching as a tool for numerical linear algebra," *Found. Trends Theor. Comput. Sci.*, vol. 10, no. 1–2, pp. 1–157, Oct. 2014.
- [38] A. C. Gilbert, J. Y. Park, and M. B. Wakin, "Sketched SVD: Recovering spectral features from compressive measurements," *arXiv preprint:1211.0361*, 2012.
- [39] A. Klenke and L. Mattner, "Stochastic ordering of classical discrete distributions," *Advances in Applied probability*, vol. 42, no. 2, pp. 392–410, 2010.
- [40] I. Johnstone, "Chi-square oracle inequalities," *Lecture Notes-Monograph Series*, pp. 399–418, 2001.
- [41] C. McDiarmid, "Concentration," in *Probabilistic methods for algorithmic discrete mathematics*, pp. 195–248, Springer, 1998.
- [42] A. Krishnamurthy and A. Singh, "On the power of adaptivity in matrix completion and approximation," *arXiv preprint:1407.3619*, 2014.
- [43] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.
- [44] G. Stewart, "Perturbation theory for the singular value decomposition," 1998.
- [45] B. Li, W. Li, and L. Cui, "New bounds for perturbation of the orthogonal projection," *Calcolo*, vol. 50, no. 2, pp. 69–78, 2013.